Data to Information to Text Summaries of Financial Data

By

Student: Cameron Kyle
KYLCAM001

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN
In partial fulfilment of the requirements for the degree

M Sc Information Technology

Faculty of Sciences

UNIVERSITY OF CAPE TOWN

18 February 2018
Supervisor: Maria Keet

Cameron Kyle, Computer Science Department, University of Cape Town

DECLARATION

I, Cameron Kyle, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 18 February 2018

Abstract

The field of auditing is becoming increasingly dependent on information technology as auditors are forced to follow the increasingly complex information processing of their clients. There exists a need for a system that can convert vast quantities of data generated by existing systems and data analytics techniques, into usable information and then into a format that is easy for someone not trained in data analytics to understand. This is possible through Natural Language Generation (NLG). The field of auditing has not previously been applied to this pipeline. This research looks at the auditing of Investment Fund Management, of which a specific procedure is the comparison of two time series (one of the fund being tested and another of the benchmark it is supposed to follow) to identify potential misstatements in the investment fund. We solve this problem through a combination of incremental innovations on existing techniques in the text planning stage as well as pre-NLG processing steps, with effective leveraging of accepted sentence planning and realisation techniques. Additionally, fuzzy logic is used to provide a more human decision system. This allows the system to transform data into information and then into text. This has been evaluated by experts and achieved positive results with regard to audit impact, readability and understandability, while falling slightly short of the stated accuracy targets. These preliminary results are positive in general and are therefore encouraging for further development.

Contents

1 Introduction

The field of auditing is becoming increasingly dependent on information technology as auditors are forced to follow the information processing of their clients. Technology such as Electronic Data Interchange enables businesses to communicate with no paper or even email trail, making it increasingly difficult to obtain sufficient audit evidence in the traditional form. Auditors are therefore forced to look towards the reliability of the client's computer and application controls as well as performing analytics on the large amounts of data generated by these applications. The availability of all financial information in an electronic format allows for new opportunities since, where previously an auditor would have to trace a paper (or PDF) invoice to a line out of the general ledger, the audit support for a particular item in an entity's financial system is another line in a separate system. This slowly led to auditors using simple excel based lookups and comparisons to obtain sufficient audit evidence, by comparing system outputs to detect differences/issues. In addition the quality and depth of financial information stored on databases has increased to the point where it is practical to use computers to perform data analytics to identify unusual trends or other patterns that would be inconsistent with an auditor's understanding of a system. This presents a problem however, as auditors are not traditionally well-versed in data analytics and are therefore not well-equipped to handle the large volumes of information that can be generated and are susceptible to being overwhelmed by information, thereby potentially missing important insights and trends. There exists a need for a system that can convert vast quantities of data generated by existing systems and data analytics techniques, into usable information and then into a format that is easy for someone not trained in data analytics to process. To focus the scope of the research, the problem of analysing the performance of a mutual investment fund against a determined benchmark of indices over a financial year was considered. Benchmarks use one or more indices (determined by the kind of assets the fund holds) to determine a market price, the change in which is compared to that of the fund. The principle behind this is that two portfolios with assets of a similar nature should respond to market movements in an almost identical way. For example, an investment fund that invests 50% in bonds and 50% in equities should have a benchmark that is equally balanced between bonds and equity. Performing this comparison would allow an auditor to identify indicators (once-off or systematic deviations in the two prices) that the overall Net Asset Value (NAV) of the fund, which is the key figure that investors are interested in, has potentially been misstated during the financial year.

As an example consider the Vanguard Balanced Fund, a mutual fund with publicly available daily price data which is benchmarked (i.e. performance measured) against a combination of the Centre for Research in Security Prices (CRSP) U.S. Total Market Index and the Bloomberg Barclays U.S. Aggregate Float Adjusted Index. For illustration a graph showing the price development of the fund during the 2016 calendar year is shown in Figure 1, which in this case shows the price of the fund matching the development of the benchmark almost perfectly. From this comparison of price data we can also derive metrics such as

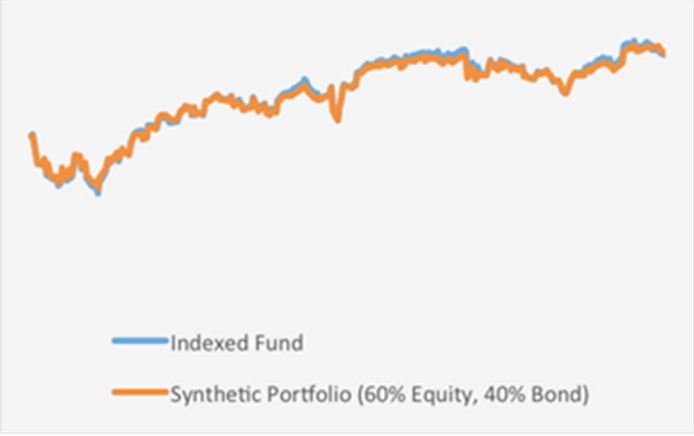


Figure 1: Illustrative graph showing the price of the Vanguard Balanced Fund as well as the benchmark (synthetic portfolio).

correlation and deviation (daily, average, maximum etc.), which can help determine an overall assessment of how well the fund is following the index. In this case an auditor would likely conclude that the fund is closely following the index and therefore that there is no indication of a misstatement.

In less clear examples (where the fund perhaps is not following the benchmark as accurately), it may be useful to supply the auditor with a written summary of the comparison instead of simply just the graph as such summaries have been shown to have advantages over pure graphical illustrations (Portet et al., 2009). Providing a textual summary to summarise the comparison between the two time series requires specific challenges to be overcome, such as how to process the data in order to make high level assessments as well as deciding the necessary level of detail to communicate to the end user.

This research aims to solve these computational and realisation challenges based largely on application of previous literature to specific problem areas into a design and to apply this design to real world data, assessing the results to determine if improvements can be made to the audit process as a result. The work of (Kacprzyk, Wilbik & Zadrozny, 2006), which applies (Sklansky & Gonzalez, 1980)'s trend approximation technique to financial time series, informs part of the early computational stage. Additionally, the fundamental work on fuzzy logic of (Zadeh, 1965) with specific application to NLG by (Ramos-Soto et al., 2015) will provide the basis from a fuzzy logic perspective. Finally, The work of (Reiter & Dale, 1997) forms the basis from an Natural Language Generation (NLG) perspective.

The literature review identifies key challenges which need to be applied to the research which are shown in Table 1. There is a significant amount of data to information transformation required, only after which the information can be converted into text. Temporal information also needs to be accurately described, which has unique challenges in aggregation, specifically how and when to combine temporal events. Finally, the extent to which full language grammar is required in comparison to summary information needs to be understood and implemented.

This research aim is achieved through a combination of incremental innovations on existing techniques in the text planning stage, as well as pre-NLG processing steps (trend identification, classification and aggregation), with effective leveraging of accepted sentence planning and realisation techniques. Additionally, fuzzy logic is used to provide a more human decision system. This allows the system to transform data into information and then into text.

Expert evaluation showed that the design produces readable and understandable output text which has a positive impact, both audit quality and audit efficiency. Further work will be necessary to improve the accuracy however, which fell slightly short of the stated target.

The research begins with a review of current literature to determine which techniques currently exist that can be applied to the problem at hand in Section 2, following which a basic design is created in Section 3. From this point, more focused research is conducted to answer specific design questions and refine the design to where it can be implemented in Sections 4 to 7. Following implementation, an assessment is performed using expert participants to provide an evaluation of the results of the design in Section 8. Finally the results of the research are analysed in Section 9 and a conclusion drawn in Section 10 on the accuracy, understandability, readability and audit impact of the design.

Table 1: A list of the challenges and solutions identified through literature review which are to be further developed in the design

Criteria	Description	Tasks	Challenge
Degree of vagueness required in data transformation	The degree to which data must be transformed into human-understandable information and the vagueness needed for such transformation can affect the techniques needed to effectively transform such data	Content Determination	Vagueness required in output information
Temporal or static information	Temporal Data would require a specific style of sentence aggregation using words such as "followed by" "dropping to X later in the day"	Content Determination, Sentence Aggregation	Continuous/ Signal Input Data
			Time series data
			Limited Corpus Available (no input data included in output)
Structure/format of intended output	Certain NLG tasks only need to output very simple pieces of text whereas others need to generate fully formed text following full language grammar	Pervasive	Developing the most effective system depending on required output

2 Literature Review

The literature review begins with a general assessment of NLG as well as presenting the framework around which further literature will be reviewed. Following that, specific aspects of NLG, fuzzy logic and trend detection which may be relevant to the research problem are assessed against the framework. The framework is then assessed against the current research problem. Finally, similar works are reviewed to determine their impact on the current research problem.

2.1 Matrix Framework

Every domain presents a different set of challenges along with solutions to these challenges, usually as a result of the nature of the input data and also the intended audience. Much of the current literature often has at its core a key problem specific to the domain as well as a solution used to solve the problem. This modular nature allows us to adopt a simple approach to synthesising the literature into a form useful for our purposes by extracting these problems/solutions in the following manner:

1. Develop a set of criteria/attributes that can be used to compare the specific challenges faced by NLG in different domains
2. Fit the various domain-based issues encountered to this set of criteria to create the desired problem/solution matrix
3. Determine the techniques used to respond to each of the challenges noted above

As an arbitrary example, we may determine that a criteria to be used in our matrix is the ratio of discreet versus continuous data (step 1). We may then determine that based on the research already done, that in the field of patient care, the data is practically continuous such as heartrate readings (step 2). Further research may then show that this has been previously handled by averaging the readings over 10 minute intervals to arrive at discrete data points in order to be converted to text.

Ultimately, we will fit the challenges/attributes of financial information to this matrix of challenges and solutions in order to determine the set of techniques that should form the optimal basis for the creation of the desired system. One potential disadvantage of this approach is that time will be spent on problems and solutions that do not apply to our problem. To mitigate this, solutions to challenges which at face value will be given little attention upfront, with the option to revisit them in detail at a later stage.

2.2 Foundations of NLG

As a starting point for this analysis we look at (Reiter & Dale, 1997) due to their formative influence. The framework laid down during the early stages of NLG research continues to be used in more contemporary pieces of literature, being cited by numerous literature pieces assessed in this paper including (Portet et al., 2009), (Adeyanju, 2012) and (Ramos-Soto et al., 2015). It combines 6 tasks into 3 stages that are systematically presented as shown in Figure 2. In the text planning phase, the message to be communicated to the reader must be identified by selecting from relevant data available (Content Determination) and putting content into a structure with a logical flow (Discourse Planning).

From there, sentence planning involves combining the short messages such as phrases into full sentences (Sentence aggregation), choosing the exact words to represent the various domain entities referenced in the messages (Lexicalisation) and determining the correct referring expressions to use such as "it" (Referring Expression Generation). Finally, the grammar rules specific to the language are applied to the sentence to ensure that it is grammatically correct as well as conveying the correct meaning by removing ambiguity (Linguistic Realisation).

For the purpose of structuring our matrix, we will use the tasks described as an initial classification for the different criteria to be established. Ultimately, these tasks may be consolidated into their respective stages or removed altogether but should serve as a sufficient starting point.

In order to proceed to more specific literature, it is useful to have a basic idea of the potentially significant criteria which can be refined based on subsequent research. The rationale for this is that, should one criterion be identified in one of the last pieces of literature reviewed, it would be necessary to review all the preceding literature to ensure that relevant insight has not been missed. By establishing some initial (if ultimately irrelevant criteria), the amount of subsequent rereading required is reduced. The results of this are show on Table 2. Four initial criteria that may be useful later in our analysis have been identified. The criteria were obtained from review of (Reiter & Dale, 1997) as well as initial knowledge of the field.

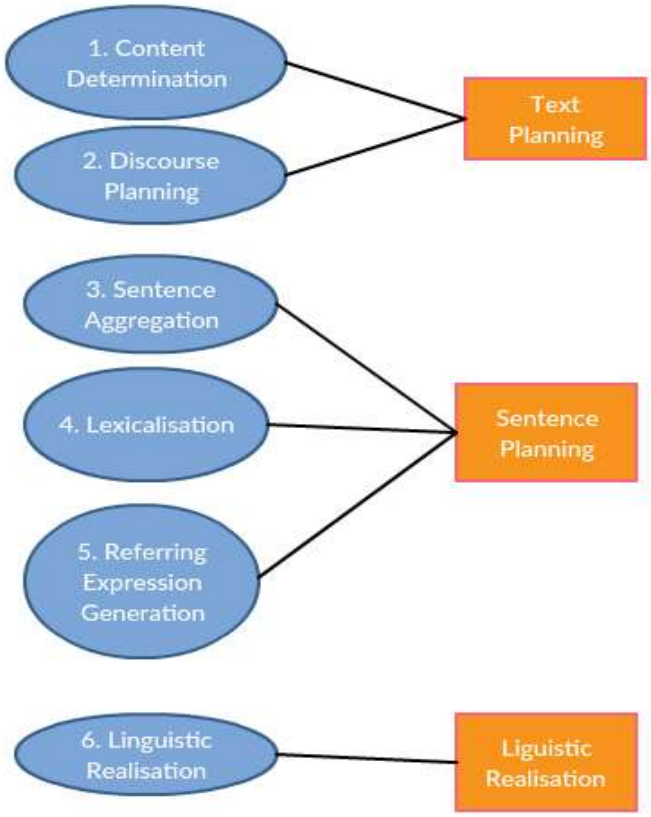


Figure 2: Summarisation of the tasks and stages proposed by [Reiter and Dale. 1997].

Table 2: Initial proposed matrix of criteria linked to NLG tasks

Task	Criteria	Description
Sentence Aggregation	Temporal or static information	Temporal Data would require a specific style of sentence aggregation using words such as "followed by" "dropping to X later in the day"
Lexicalisation	Variability of user's ability	If the text is to be presented to users with varying reading ability additional complexity is introduced
Linguistic realisation	Availability of corpus	The availability of an extensive corpus can influence the techniques available to a designer
Pervasive	Structure/format of intended output	Certain NLG tasks only need to output very simple pieces of text whereas others need to generate fully formed text following full language grammar

2.3 Temporal vs Static Input

As a starting point, the first criteria will be used for further research by analysing literature that focuses on domains with temporal data to determine if there are challenges particular to that domain. In their description of Babytalk, a system which generates textual summaries of multiple vitality readings in neo-natal intensive care, (Portet et al., 2009) explored problems specific to time-series data. One such problem is how to handle the large quantity of information contained in practically continuous dataset (1 sample per second over 45 minutes totalling 2700 total data points for one reading). The technique used to solve this problem was temporal abstraction which involves combining multiple sequences of events into larger patterns. To do this, small shapes (such as spikes in a reading) are defined in the system's ontology, and from there larger patterns are defined as specific combinations of particular shapes. The first problem (continuous data) and solution (temporal abstraction) to a previously-identified criteria (temporal data) has therefore been identified.

As for the process of actually identifying trends, significant research has been performed on the subject in the realm of data mining, a review and summary of which is given by (Fu, 2011). Various methods are explained, ranging from a simple sampling approach (dividing the time series in separate pieces based on a pre-determined sampling interval), to more advanced approaches such as Symbolic Approach Approximation (SAX). One version of this approach is presented by (Lin et al., 2003), in which a time series is first broken into samples where only the mean value across the sample is retained. The various samples are then given alphabetic classifications (e.g. "a, b or c") corresponding to where in the possible range of values the mean for the particular sample lies. The result is a sequence of letters (e.g. "baacccdddabc") which describes the time series in symbolic terms and allows for comparison to other time series as a way to efficiently process the time series in data mining. This is unlikely to be a useful approach in the current problem, since the process of taking average values for the sampled sections removes the possibility to accurately tell where an identified trend begins/ends. In addition, the result is a representation of the various levels of the time series and says nothing about the trends indicated by the time series (this would still need to be determined based on the various levels).

Another approach, presented by (Keogh & Pazzani, 1998), involves the weighting of certain parts of the time series more than others, based on how much information is contributes to the overall summary. This was used in stating how similar two times series are, but is not able to describe the time series itself. The lack of suitability to the current problem is likely a result of the different objectives in the data mining domain. A key objective in data mining is space efficiency (Lin et al., 2003), both in storing data as well as processing it, as data mining is processor-intensive. This makes it useful to have representations such as ("abbc") for a time series. In the current problem however, there is no such constraint on space, however there is a requirement to be able to describe the time series and not just conclude if they are similar.

An alternative approach is described by (Kacprzyk, Wilbik & Zadrozny, 2006) based on the work of (Sklansky & Gonzalez, 1980) which analyses the time series in a piecewise manner. From the starting point in the analysis (the first point which does not fit the previous trend) cones are drawn for each subsequent point which are tangential to circles drawn around each subsequent point with radius equal

to a predetermined tolerance measure. This can be seen graphically in Figure 3. The possible trend area is determined as the area in which all sequential cones intersect, with the trend ending with a point whose cone does not intersect the previous ones. One major benefit from this approach is that the computation only needs one pass on the data set (Kacprzyk, Wilbik & Zadrozny, 2006) since there is no requirement for retroactive adjustment of the trend.

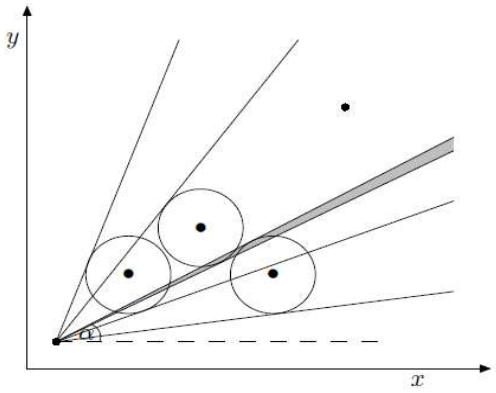


Figure 3: An illustration from [Kacprzyk, et al. 2006] showing how cones are built from a starting point which determine a trend at the intersection of subsequent cones.

(Portet et al., 2009) further go on to describe another problem caused by a combination of temporal data and the ordering of text based on importance. Their Babytalk system, BT-45 ranks messages based on estimated importance and then attempts to present these messages in order of importance. This can lead to misinformation if an important but late-occurring event is presented first without careful consideration of the syntax used to relate it to subsequent messages. As an example, if the text presents a large spike in the patient's heartrate and immediately follows with a downward oxygen trend (which in fact occurred prior to the spike), the reader may infer that the heartrate spike caused the drop in oxygen levels. This is compensated for by grouping messages into a paragraph with one root (key) message with the start time for this message being explicitly stated each time. This allows readers to orientate themselves when considering the order of root events.

2.4 Availability of a Corpus

The description of the Babytalk system has highlighted two challenges and solutions specific to temporal data. The use of this system however was based on a neatly constructed corpus of instrument readings (input data) as well as human compiled output directly related to that data. This allows for a wider range of options with respect to developing, as well as teaching, a system.

> One such example is CBR-METEO, a weather forecast NLG system, in which its creator (Adeyanju, 2012) describes the use of case-based reasoning (CBR) applied to NLG to utilise a library of pairs of input data and human generated output (a case) to convert new input data into output text. Practically the system follows four steps (Retrieve, Reuse, Revise, Retain) to select the most relevant case to the input data, refine the case's output to suit the specific input data and present this to the user.

As noted by (Adeyanju, 2012), the combination of CBR and weather forecasting NLG is highly effective since the components in a weather forecast are often repeated (e.g., wind speed: it can increase, decrease or remain the same) and therefore a corpus could easily be grouped into a few categories of inputs, each of which has multiple outputs associated to it. This allows for more options to improve the system's output variability. We can conclude that when a corpus is available, which contains both input and output data, and the input data is repetitive, CBR can be a powerful tool in building an NLG system. This is not a challenge/solution pair but rather an opportunity/exploitation. Nevertheless, it is useful to include in our matrix.

While teaching a system based on corpora may seem promising, (Reiter & Sripada, 2002) caution against using it without carefully considering potential deficiencies in the corpora. They ultimately warn that corpora-based learning requires sufficient consideration with regard to the quality of the corpora being used as well as the potential for pervasive errors or inconsistencies contained within.

What would be the options available if no corpus is available or available in a limited format? (Reiter & Dale, 1997) describe a corpus as a collection of inputs and their associated outputs consistent with our use of the term so far. (Smiley et al., 2016) however use the term slightly differently, since as their corpus consists purely of examples of output text in a particular domain, in this case Reuter's news articles, which have no corresponding input. For the sake of disambiguation we will classify this as a limited corpus. (Smiley et al., 2016) present a technique for verb selection based on analysis of the frequency of verbs used in conjunction with numeric changes. This is illustrated in Figure 4. Based on this data extraction, a statistical analysis is performed to determine which pairs of input (PercentMove) and output (verb) phrases occurred most frequently. Using these frequency distributions, the most appropriate words for certain percentage movements (separated between upwards and downwards movements) could be determined and applied to the linguistic realiser. Ultimately, the result is more natural sounding and meaningful text in place of magnitude-independent synonyms for increase/decrease. This technique effectively allows a developer to utilise a limited corpus and transform it (at least for some intents and purposes) into a full corpus, containing both input (percentage movement) and text output (verb describing the movement).

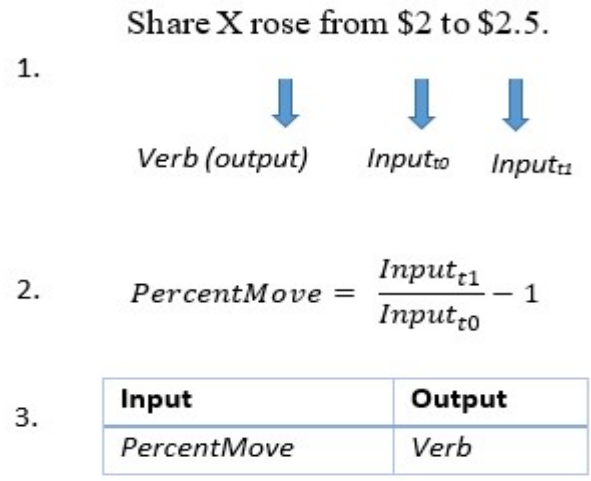


Figure 4: Explanation of the process to convert a limited corpus to a full corpus.

In the case of a limited corpus which lacks sufficient information to extract input data, there are still techniques available to generate natural language. One such approach is described by (Knight & Hatzivassiloglou, 1995), which uses a large scale and even domain independent corpus, in this case a collection of Wall Street Journal texts, to determine n-gram probabilities. It uses these probabilities to filter out many different possible sentences generated by the module(s) in the sentence planning phase. The key impact of this approach is that the sentence planning phase does not need to be overly refined by manually encoding rules to deal with ambiguity or certain types of syntactic agreement that a system would otherwise not be able to handle programmatically. The system was found to be particularly effective at handling grammatical cases (e.g., "she"/"her") as well as possessive pronouns. It was not necessary for the grammar engine to distinguish the correct case or distinguish a possessive pronoun from an adjective, as incorrect options would be filtered out. The use of relatively small n-grams ($0<n<4$) therefore provides a very effective filter for short-distance agreement (words sequentially close in a sentence) since both the words required to be in agreement will fall into one n-gram, such as "his head" in a bigram.

What (Knight & Hatzivassiloglou, 1995)'s model was not able to do, was handle longer-distance agreement, since the affected words would not appear in the same n-gram. They alleviate the problem through the use of many paths generation, also known as word lattices. These work by allowing the

system to choose between multiple paths, each of which only contain word options which in any combination will be in agreement.

> For example the phrase “A big golden retrievers” would not be prevented by any n-gram below n=4 however a word lattice could be constructed where “A” is in a path without plural nouns.

It should be noted that the authors’ objective was to design a language translation system which meant the grammar had to be context free. That is not the case in the majority of NLG literature described in this review, where there is a specific domain in which a system has to function.

It is also useful to consider some existing options for ready-made domain independent grammar engines, as these require far less effort than implementing the algorithms mentioned above. SimpleNLG is perhaps one of the best-known grammar engines and is described by its authors in (Gatt & Reiter, 2009). It is designed to be efficient, robust and set up to allow for flexible method of inputs and while it does not have the most comprehensive coverage, it covers most common English words and can be extended via an additional lexicon. For more complete coverage there is the Komet-Penman Multilingual-Lingual (KMPL) which has a larger (and multilingual) lexicon and can be run as a black box server, accepting semantic input and providing text output (Bateman, 2016).

2.5 Domain Independence

This above problem/solution illustrates another potential criteria which is the degree to which the system has to be domain independent. Adjusting the rules and vocabulary to a domain may result in 'overfitting' and thus less domain independence. On the other hand too generic a rule base may result in missing the needed specialist vocabulary or jargon for the specialist in the domain. The problem in the example sentence above could be avoided by simply limiting the number of descriptors between the referring expression (“A”) and the single/plural noun (retriever), allowing small n-grams to filter out disagreements. This limits the flexibility of the system, but in a domain-specific system this may be acceptable. For example, in a weather forecasting system such as the Forecast Generator (FOG) system described by (Goldberg, Driedger & Kittredge, 1994), the range of nouns that exist in the domain is limited to meteorological concepts such as wind speed, wind direction and rain fall. Additionally, the system only needed to describe the initial state and changes to those concepts. This makes it more feasible to plan the types of agreement which need to be enforced. The FOG system would therefore be on the opposite side of the domain-dependence spectrum compared to a language translator. This contrast presents another criteria to add to our matrix, the extent to which the system needs to be domain-independent. Systems that do not need to be domain-independent can utilise many paths generation (word lattices) in combination with n-gram filtering to produce high quality text that is free from most forms of disagreement.

2.6 Complexity of output

With regard to the variability of the users reading ability, (Moraes, McCoy & Carberry, 2016) describe an approach to vary sentence aggregation based on the requirements of the intended user. They argue that textual summaries are most effective when tailored specifically for the reading ability of a target reader. The system starts with all intended message separate as the initial state and from there combines

the messages, meanwhile computing a cost function for the marginal complexity added to the text by each successive message to a sentence. The complexity cost is based on 15 metrics such as average sentence length and percentage of adverbs. The system was subjected to multiple evaluations in which it performed well, being able to achieve statistically significant results. The challenge (how to customize readability to the needs of the user) is therefore effectively handled through this method.

Regarding the complexity of the output, (Reiter & Dale, 1997) note that something as simple as a mail-merge functionality is a basic form of NLG, an extreme example of a template-based approach where the entire message's structure is predefined and only key variables defined. On the other end of spectrum the NLG system could be free to determine the sentence and even paragraph structure without human intervention (often referred to as "real" NLG systems (Van Deemter, Theune & Krahmer, 2005)). This provides the advantage of more varied and less monotonous text with the increased risk of generating grammatically incorrect text, since it is difficult to program all the necessary grammatical rules. Numerous articles discuss the differences and advantages of each approach. In one such example, (Ramos-Soto et al., 2015) provide a good example of utilising each approach to tackle different challenges within the same NLG system. Their system Galiweather (forecast summary generator for municapilities in Galicia) utilises a template approach for the simpler variables encountered in weather forecast generation, such as wind speed and direction, which do not change more than once or twice during a forecast cycle. A more flexible sentence constructor is built for variables such as precipitation which can change multiple times throughout the forecast period. In implementing this second component three sub-modules were created to handle the aggregation of basic messages created in the content determination module, each with a different level of aggregation (episode, day and forecast term). Each module also performs the remaining NLG tasks, such that the output is three fully formed sentences independent from each other, from which the shortest (assumed to be the simplest) is chosen as the output of the precipitation component. Like (Knight & Hatzivassiloglou, 1995)'s work described above, this is another example of using a computer's efficiency to quickly perform semi-redundant tasks which produces multiple candidate outputs, of which the optimal output can be selected. Looking to our identified criteria, it can be seen that this challenge and solution pair relates to the structure and intended format of output criteria that we previously identified. Simpler texts (e.g. conveying wind speed/direction) were created using templates while more complex formats (precipitation) were created using "real" NLG techniques.

2.7 Fuzzy Sets

Another interesting aspect of the Galiweather system described in the preceeding sub-section is the application of fuzzy logic to the content determination stage. Fuzzy logic, and more specifically fuzzy sets, differ from the classic crisp sets in that membership for a fuzzy set can take on a value between 0 and 1 (unlike a crisp set whose membership is either 0 or 1) (Zadeh, 1965). This approach allows more complicated and less clear-cut (hence fuzzy) systems or problems to be formally modelled and computed. In the Galiweather system, this is applied to temporal variables, among others, which can be seen in Figure 5.

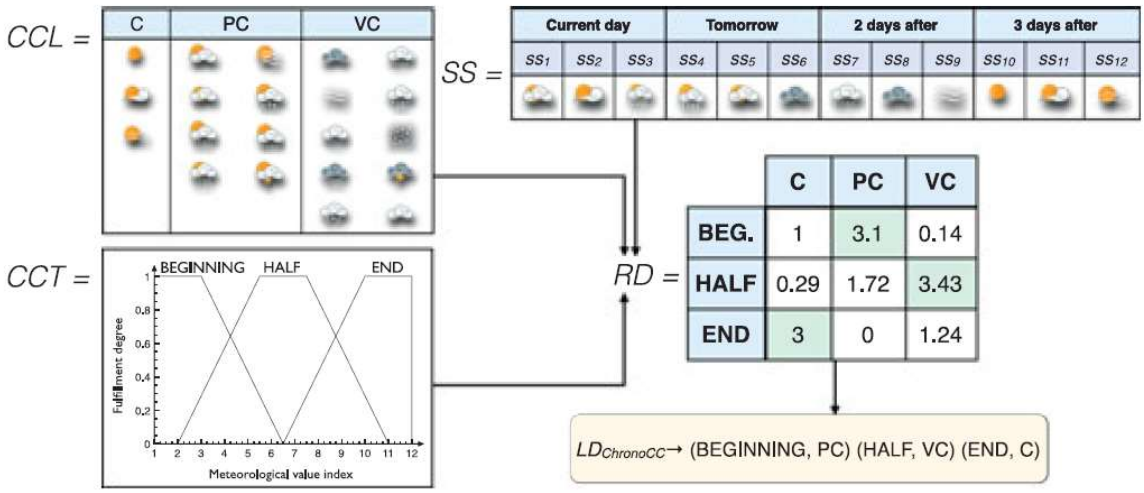


Figure 5: An extract from (Ramos-Soto, et al., 2015) which illustrates the use of both fuzzy and crisp sets to determine the descriptors for weather forecasts. The CCL table shows the possible Cloud Cover Linguistic variables. CCT shows the Cloud Cover partitions. The SS tables indicate the various Sky States. This is combined into a Relevance Degree (RD) matrix.

The forecast term used in this system is a few days with 12 distinct forecast intervals, each with a different cloud cover variable (a crisp set). Fuzzy sets were created for the beginning, half and end of the forecast term. Based on the membership functions, the membership function for the third interval was 1 for "Beginning" but only 0.25 for "Half". The membership scores on the crisp sets (cloud cover) and fuzzy sets (term) are combined to determine the most relevant cloud cover set per term. The practical impact of the fuzzy set is one interval's cloud cover can influence the overall score for more than one term (i.e. it can give a greater score to "cloudy" in both the beginning and half terms) which would not be possible with crisps sets alone. It further means that when transforming data into information at the content determination stage, there is no need to specify an exact classification for each value a variable may take on. Instead the classification can be made more flexible which is more in line with human thinking and behaviour (Zadeh, 1973), a highly desirable ability in a system which needs to produce human-like output.

From the above approach we have identified another criteria, challenge and solution to add to the matrix. The criteria is the degree to which data (temporal intervals in this case) must be transformed into more natural information (beginning, half and end of the forrecast period) and the vagueness of such transformation. The challenge is how to handle such transformations, since the vagueness is significant. Lastly, the solution is the use of fuzzy sets to allow for the vagueness to be maintained while also allowing for further computational processing.

2.8 Final Matrix

Prior to finalisation and as indicated at the outset, all criteria identified subsequent (Degree of vagueness and domain-independence) to the initial matrix (under subheading "Foundations of NLG") must be re-assessed. This is to ensure that literature assessed earlier, which has relevance to the the criteria, challenges and solutions identified later, is not ignored.

Regarding the last criteria (vagueness of information transformation), it can be seen that there is similarity between the the temporal abstraction performed by the Babytalk system and the use of fuzzy sets utilised by Galiweather in that both are involved in the content determination stage and both handle temporal data. Looking closer, it can be noted that these two techniques function at slightly different stages of the same task, specifically aggregation of input data and fuzzification of the resulting information. Temporal abstraction takes signal input and converts it into discrete intervals while the fuzzy sets in the Galiweather system took the intervals and gave them membership scores for each term. There may be a possibility to bypass the abstraction stage and apply the fuzzy sets to signal data (e.g., by determining the start and end of "episodes" as the point the signal starts leaving one fuzzy set or starts entering another). This is beyond the scope of this research however.

Aside from this possible interaction, there is no other impact of the previous pieces of literature. The review is threfore concluded with the matrix shown in Table 3, first 5 columns.

2.9 Application of the Matrix

To use the matrix, it is first necessary to assess the specific problem area against the criteria identified in the matrix to determine which techniques may be of use in solving our problem. This is shown in Table 3. As originally stated, the goal is to create a system that can convert vast quantities of facts learned from financial information (generated by existing systems and data analytics techniques) into a format that is easy for someone not trained in data analytics to comprehend and use. In order to focus the research, we will concentrate on one particular task that an auditor may need to perform, a trend analysis of financial time series information, specifically a comparison of the performance of a mutual fund to its intended benchmark. The variation of the Net Asset Value (NAV) of a fund to its intended benchmark can indicate the existence of errors or fraud within the mutual fund. An auditor can therefore compare the NAV over the financial period of the mutual fund to the price change of the benchmark(s) over the period. Through inspection of the two time-series, areas of risks can be identified (refer Figure 6 for an example) and from there more specific audit procedures can be performed to determine if a misstatement exists. The intended system therefore needs to be able to take two time series as inputs and provide as output a textual summary of the degree of similarity (in human terms, not simply starting statistics), a concise summary of where deviations were noted and potential reasons for the deviations identified.

As stated the system has to take time series as data input and provide output in human terms (linguistic descriptions). The vagueness of data (one cannot draw a simple computational conclusion from two time series) along with fuzzy sets are therefore applicable to our system. While not on the same scale as the readings used by BT-45 the system still needs to be able to handle roughly 262 (working) days' worth of data points. At the same time the objective does not require summarising only one time series but two. There is therefore less importance in determining periods of interest based only on one dataset but rather based on the relative change between two time series. This may afford new opportunities and techniques in content determination with regard to temporal abstraction. For the same reason the ability to identify trends using linear approximation of cone-graphs is also relevant to the problem.

Table 3: Final matrix with challenges and solutions encountered so far. Two criteria have been added subsequent to the initial matrix with a total of 9 challenges/solutions identified as well as the applicability to the current problem

Criteria	Description	Tasks	Challenge	Solutions	Applicable?
Degree of vagueness required in data transformation	The degree to which data must be transformed into human-understandable information and the vagueness needed for such transformation can affect the techniques needed to effectively transform such data	Content Determination	Vagueness required in output information	Fuzzy sets	Yes
Temporal or static information	Temporal Data would require a specific style of sentence aggregation using words such as "followed by" "dropping to X later in the day"	Content Determination, Sentence Aggregation	Continuous/ Signal Input Data	Temporal Abstraction	Yes, with modification
			Time series data	Trend detection using linear approximation based on cone graphs	Yes
		Sentence Aggregation	Information Rank vs temporal order conflict	Paragraph grouping with one key event limit	No
Variability of user's ability	If the text is to be presented to users with varying reading ability additional complexity is introduced	Sentence Aggregation, Lexicalisation	Ensuring readability on a user by user basis	stepwise sentence aggregation with marginal complexity cost analysis	No
Availability of a corpus	The availability of an extensive corpus can influence the techniques available to a designer	Linguistic Realisation, Lexicalisation	Full Corpus Available	CBR applied to NLG	No
			Limited Corpus Available (input data included in output)	PoS tagging with Frequency distributions	No
			Limited Corpus Available (no input data included in output)	Multiple sentence generation refined by n-gram analysis	Yes
Structure/format of intended output	Certain NLG tasks only need to output very simple pieces of text whereas others need to generate fully formed text following full language grammar	Pervasive	Developing the most effective system depending on required output	Use of templates vs “real” NLG techniques	Yes
Domain-Independence	The extent to which a grammar has to be context-free can influence how robust the system's grammar engine needs to be	Referring Expression generation, Linguistic Realisation	Require Context-Free Grammar	Many Paths Generation (Word Lattices) or packaged NLG software	No

Regarding the temporal order, although the proposed system will handle temporal information, the importance of the temporal order of information conveyed is not as high a priority. In the BT-45 system, high importance was placed in ensuring that the reader understood the exact sequence of events coming from different readings described in the summary (since a user’s assessment of causality could be affected by that understanding). In the proposed system, the user is more interested the presence of deviations and not the order they occur in. Furthermore, since there is only one metric of interest (price at time x) there is no cross causality, such as that noted in the BT-45 environment.

The users of the intended system would have an expected level of expertise in the field of auditing implying a relatively high level of education and linguistic ability. The system would therefore not

need to cater for a wide range of reading abilities therefore eliminating for techniques such as marginal complexity cost aggregation.

On the subject of corpora, to the best of the author's knowledge, there is no available corpus containing comparative time series as well as human generated textual summaries of said time series. This limits the applicability of CBR to the current problem. Additionally, since the input primarily consists of two time series it is not possible that this input could be included in output text, as in the share price example provided in Subsection 2.4. While there is no corpus directly related to time series comparison, there are corpora covering business related language which could provide a useful basis for simple n-gram analysis to provide the system with the linguistic ability for communication in a professional business environment such as the Reuter's Corpus (Lewis et al., 2004). Such a corpus may prove useful in determining (at minimum) appropriate descriptive synonyms to be used in the system.

Lastly, the intended output of this system is a description of the differences and similarities between two time series, with consistent metrics on each axis. It is therefore clear that the output (while attempting to avoid monotony) will have certain repeated elements, which make a template-focused approach feasible rather than trying to build entire sentence structures for each instance of the output. This also indicates that the system will not need to be domain independent, limiting the usefulness of many paths generation. For simpler sentences however, it could be useful to use packaged NLG software which should provide realisation with a high degree of reliability.

2.10 Similar Work

Going further from general NLG research to literature more focused on the specific problem set, there is one major series of work which tackles the same problem of comparing time series relating to mutual funds. (Kacprzyk & Wilbik, 2009) first tackle the problem of time series comparison for a mutual fund and its benchmark by building on several of their previous papers, which consider the application of fuzzy logic to the description of time series. In this work, they utilise the trend determination based on linear approximation of cone graphs technique (previously discussed) to identify separate trends for which textual summaries are generated based on protoforms as described by (Zadeh, 2002). Protoforms use simple linguistic sentences such as "most pets are brown" where "most" is a fuzzy quantifier and "brown"/"pets" are fuzzy labels which together represent knowledge in a form that is understandable to humans as well as being computable. In (Kacprzyk & Wilbik, 2010), they use protoforms to generate separate descriptive sentences for each time series based on the duration, dynamics and variability of the trends previously identified and make overall statements about the nature of the trends.

> As an example the sentence "among all y, most are constant" was generated meaning that the majority of the trends in the time series had a fuzzy label for the dynamics measure of "constant". Each of these statements has a truth score attached based on (Zadeh, 1997)'s calculus of linguistic quantified propositions which considers the weighted average of trend's membership score based on membership functions of the summarizer ("brown" in our example) the value of which is finally applied to the membership function of the linguistic quantifier ("most"). These overall statements were then compared between the time series and a similarity score calculated which is based on the similarity between all the statements which had truth scores above 0.8 for either time series.

While the above is highly summarised, the key point is the overall philosophy behind the approach proposed which is to consider each time series in isolation and generate descriptions that best describe it. Only after this is done for both time series are the results compared to determine the how well each time series match the other.

While the work presented above showed promising results from a technical standpoint, it is highly focused on the content determination stage of the NLG problem and does not address the issue of actually creating a full NLG system with planning and realisation missing. This means that an approach is still required for the entire problem (though it does provide some useful foundations for further content determination).

Additionally, while this is an interesting application of fuzzy logic and provides potential for multiple other avenues of time series comparisons it is not well-suited for our particular problem. (Kacprzyk & Wilbik, 2010) sought to provide a measure of how well a fund performed compared to its benchmark in order to establish if the fund was providing the return that it should be and therefore whether it is likely to perform consistently in future. From an auditor's point of view however the fund's past or future performance is irrelevant as long as it is accurately represented i.e. an auditor is indifferent between a fund that loses 20% or gains 20% in a year provided there is no evidence that the price was manipulated. What an auditor does care about however is indications of possible tampering. As a basic example consider Figure 6 which shows otherwise identical time series except for one large deviation. Based on the above method, a description "most trends are constant" would achieve a very high truth score in both time series since the one large deviation would be lost in multiple other conforming statements (this is intuitive since it is hard to argue against the assertion that most, but not all, trends in this graph are constant). These two time series would therefore achieve a very high similarity based on (Kacprzyk & Wilbik, 2010)'s approach, ignoring a large deviation which to an auditor could be very significant.

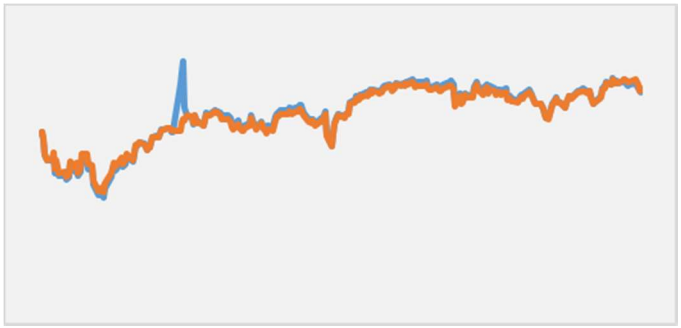


Figure 6: Graph showing two times series which run almost in parallel except for one large deviation. Data is based on actual fund and index data but modified for illustration.

For these reasons, there is a still a need for a new and complete approach to tackling the problem of developing an NLG system to communicate the comparison of times series of mutual funds and their benchmark. A system is required which can focus on and describe the deviations in a comparison, rather than looking at the similarities.

As a result of the literature review, various techniques are available which to begin building a design. Following this literature review, further work will be required to refine these tools and apply them to this specific domain and research problem. Section 3 shows the initial design as a result of the literature review whereas sections 4 to 7 go into further detail on specific areas.

3 Initial Design and Input

3.1 System Overview

The system architecture for the design is shown in Figure 7 below and subsequently described.

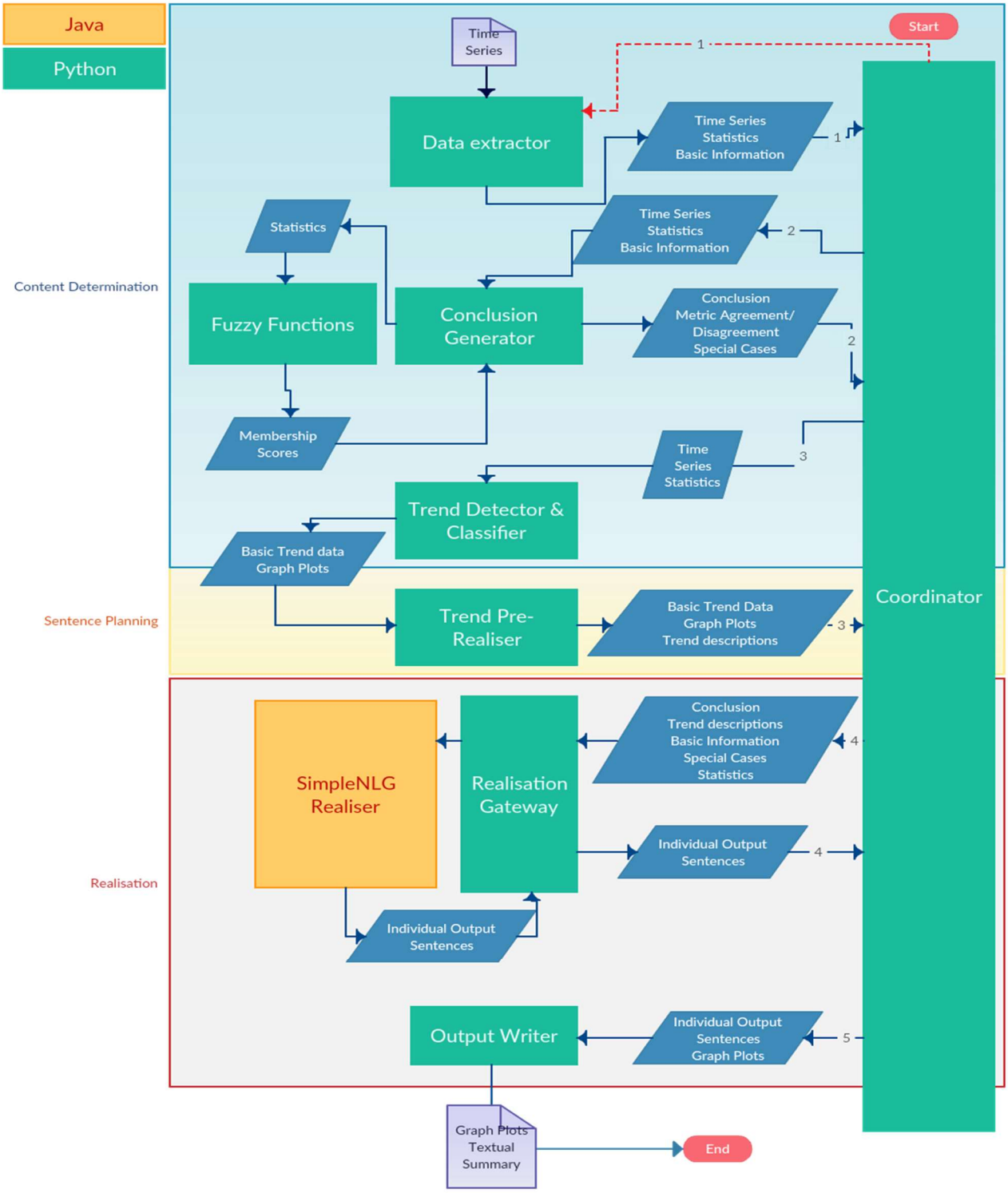


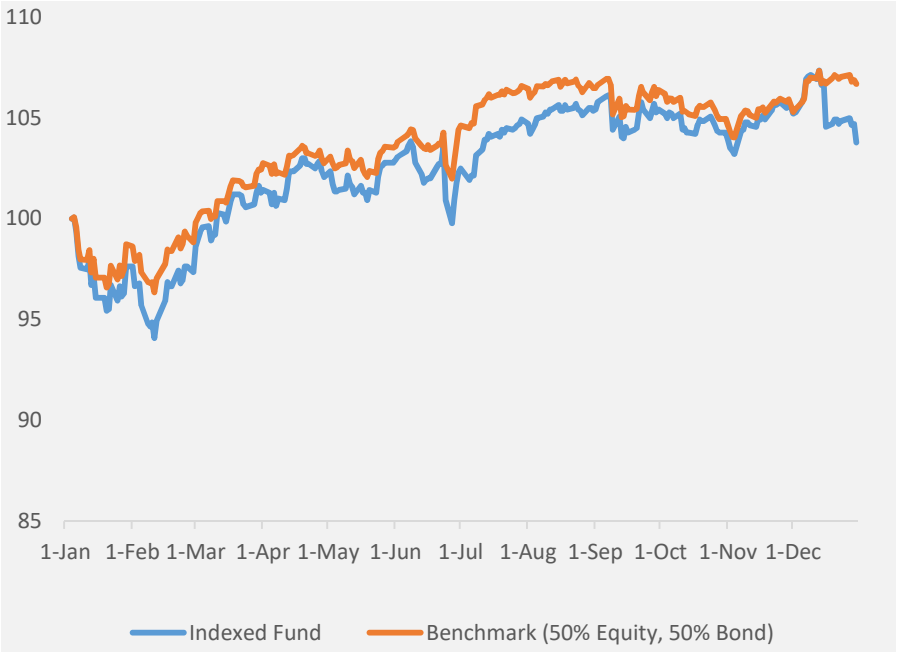
Figure 7: The system architecture for this design. Rectangles show the core modules, blue parallelograms show data flowing between them.

Data flows from the top down via the coordinator, responsible for managing the information flow, to arrive at the final summary. Content determination is the largest of these tasks since significant processing is required on the provided data to generate information. In this sense it is not a data-to-text system but rather a data-to-information followed by information-to-text design.

3.2 Content Determination

Following on from the initial literature research, we present the intended outline of the system. From this point, further research will be conducted on the challenge/solution pairs previously identified in the matrix. It is therefore important to define a clear structure for the system including input, processing and intended output. In describing the structure we make use of (Reiter & Dale, 1997)'s NLG system design architecture with some tasks being aggregated as appropriate. Note only an outline of these activities are provided here with further detail included with additional research. For implementation language Python was chosen as the implementation language simply based on author preference and experience.

The system will receive as primary input two sets of time series data representing the daily prices of a mutual fund and its comparative benchmark. An example of the input is presented in Figure 8 which shows a graph plot of the input data as well as example of a potential summary. The determination of the benchmark is made by the user based on their needs and is outside the scope of the system. The time series represents the indexed price of the fund of base \$100 with the start of the time series being the base period i.e. if the Fund's NAV at t_1 is \$221 and \$243 at t_2 then the indexed price becomes \$100. The same is done for the benchmark as this is necessary for direct comparison.



As shown above, there is relatively strong correlation between the funds of over 98% but a variation range of 0.21% to 2.81%, the latter occurring at the end of the period. The benchmark is generally higher than the indexed fund with a significant positive variation in February and from July to September where the average variance was 1.33%. An event toward the end of the period appears to have caused the maximum deviation. The average NAV was 1.07% higher for the benchmark compared to the Indexed Portfolio.

Figure 8: A graph of the two time series that will form the input data to the system along with a human-generated textual summary. In this case the Indexed Fund represents the indexed price of the JP Morgan Balanced Fund while the synthetic portfolio consists of the S&P 500 and S&P Bond Indices in equal parts over the 2016 year.

The absolute price of each investment is irrelevant since, given the same amount of money to invest, you could invest the same amount in either investment at the beginning of the period (with differing quantities of shares/units) with the only relevant factor being how much either investment would provide as return at the end of the period. Any reference to investment prices from this point on will refer to the indexed price.

From this the data the following computations will be performed:

- Basic summary statistics as well as daily measures (such as % variance) will be calculated (Section 4.1)
- Fuzzy logic will assign key statistic values to fuzzy sets which, when combined into a decision matrix, will determine an overall assessment for the similarity of the two time series. (Section 4.2)
- Key trends/events will be highlighted and described in the comparison of the time series by first identifying trends, classifying trends/groups of trends and then aggregating similar trends into one message. (Section 6)
- Additional summary information will be provided (Section 7)

3.3 Sentence planning

The overall assessment determined in the text planning stage will determine a loose template for the final output, determining what kind of information should be conveyed to the user. As a basic example, if the correlation is near 100% with little/no daily variation then there is no point in describing trends, since there will be no difference in the time series. The templates will allow for key information to be inserted into the sentences and will limit the necessity of grammar engines. A similar approach has been implemented in the evaluation of line graphs by (Moraes et al., 2014) where key measures of a single line graph determine the type of information to be conveyed in the summary. A key difference in the proposed approach is the use of fuzzy sets as a way to make the decision process more in line with human thinking.

The determination of which metrics are to be included as inputs in the final assessment will be the main focus of further research into this task. Additionally, the exact method for calculating these metrics will also be refined and explained. Following this, the method for determining which combination of sets constitutes the optimum measure also needs to be determined through further research. Lastly, the exact fuzzy functions and decision matrix need to be decided.

Following this step, the information to be conveyed to the user will be customised to fit the overall assessment. This will be implemented through a combination of canned strings as well as semantic structures that will be passed to the realisation stage. The exact distinction between canned and “real” NLG components within the text is the focus of further research. In addition the overall assessment will determine which trend information is required (if any) in the following step. A dictionary providing a mapping between assessment and trend function will allow the system to only run the necessary algorithms for each assessment scenario. Sentence aggregation will be handled in a canned manner i.e. which messages belong together will be predefined based on the template, though the form of the message could still be dynamic.

A key piece of information to be included in the final output is a description of the major trends of the graph (dependent on the overall assessment). Although a method for identifying trends has been

selected, it is considered prudent to perform a brief assessment of other possible options (of which there are many) in the field. This is to identify a potentially superior algorithm. Once the trends are identified, they will be classified again using fuzzy sets where a trend with a calculated gradient and length will be assigned a membership function (e.g. sharply increasing). Fuzzy sets are used over crisp sets for more human computation.

> Consider an example where 4 trends with largest membership scores in ("very sharply increasing") and 2 other trends all with ("increasing"). In both cases however the trends both have lower membership scores in the other gradient set ("increasing" and "very sharply increasing" respectively). In this situation, it may be more preferable to combine all trends into one category (either "increasing" or "very sharply increasing"). This category would be that with the highest combined membership score.

The details for implementing this approach will largely follow those in the overall assessment fuzzy sets however, further research will be needed to determine the exact membership functions.

Regarding referring expressions, no use of this will be made in the design as they could be confusing to the reader. Since there are two entities being described (the Fund and the benchmark), each with the same properties, using non-specific referring expressions could lead to confusion (e.g. "the fund performed better than the benchmark. Its growth was 4%") which, although generally clear, could lead a reader to assume the wrong entity is being referred to. Lexicalisation will be handled in greater detail together in sub-section 7.2.3.

3.4 Realisation

Lastly, during realisation the semantic structure obtained from the previous two steps needs to be converted into fully formed and grammatically correct sentences. To do this, SimpleNLG will be used as the realisation engine. It is preferred over KPML due to its simplicity and robustness, and since the output would not require an extensive vocabulary (given the relatively narrow focus on the characteristics of a time series comparison), any shortcomings in coverage could be manually added to the lexicon. In order to utilise SimpleNLG in conjunction with the Python language, the SimpleNLG code (Java) will include functions for each type of realisation structure required, taking strings as input, and providing strings as output to the main system, thereby negating any concerns about exchanging incompatible data types. The main research required for this task is to determine how the semantics should be structured in the content determination tasks so that they can be sent as the correct arguments and with the appropriate function to the SimpleNLG module.

The above tasks and questions are summarised in Table 4.

Table 4: The main design tasks and further questions which need to be answered as part of the more detailed research.

Focus Area	Further questions to be answered
1. Overall assessment decision using fuzzy logic	Which metrics will be used to make the assessment?; How will these metrics be defined?; How will the fuzzy functions be defined?; What will the decision matrix look like? and; On what basis will the best fitting combination of fuzzy sets be chosen?
2. Assessment – dependent template creation	What will be the exact split between canned text and "real" NLG components? What information will be provided based on the overall outcome and specific situations determined?
3. Trend identification, classification and aggregation	Is there a better approach to identifying trends? What will the membership functions for the trend metrics be and how exactly will they be combined?
4. Realisation choice: full grammar engine or many-path generation?	How will the semantic component be structured for further processing into the realisation stage?

The summary begins by stating the overall key statistics (discussed further in the next chapter) and then goes on to provide more detailed descriptions of trends and anomalies noted in the time series comparison. It then goes on to provide additional information based on the initial assessment. In this case, the average variation was provided along with a clarification as to which time series (benchmark or fund) was higher on average. It would be simple enough to simply state the deviation with a +/- sign and leave the user to determine the impact of the sign however the purpose of the textual summary is to lessen the amount of information processing required by the user. In the following chapter, it will be seen how the structure of this output can vary based on the overall assessment. In this case the structure of the text could be something along the lines of:

As shown above, there is relatively [strong]$_{computed,fuzzy}$ *correlation between the funds of* [over]$_{computedRounding}$ [98]$_{computed}$*%* [despite]$_{templateDependant}$ *a variation range of* [0.21]$_{computedMin}$*% to* [2.81]$_{computedMax}$%, the [latter]$_{highest}$ *occurring at the* [end]$_{fuzzyTimeinPeriod}$ *of the period.*

This is only an indication since ultimately the degree to which the text is "canned" i.e. fixed strings occupying a known place in the output, is still to be determined based on further research.

Based on input from other industry experts through informal discussion, insight was gathered into the additional kinds of information which would be useful for the system to output. This included providing a brief summary of the key characteristics of the fund which would give the auditor sufficient context in which to interpret the results. Such characteristics would include:

- The overall nature of the fund i.e. is it a passive fund that seeks to track an index/mix of indices or is it an active fund which seeks to outperform an index.
- The type of assets held by the fund as well as the assets it is allowed to hold

From the above analysis there are 4 areas where further research:

4. Overall conclusion using fuzzy logic
5. Template Degree
6. Trend identification, classification and aggregation
7. Realisation

These will be discussed in the following sections.

4 Overall conclusion using fuzzy logic

This section explains the further research into the first design tasks and answers the following questions:

1. Which metrics will be used to make the assessment?
2. How will these metrics be defined?
3. How will the fuzzy functions be defined?
4. What will the decision matrix look like?
5. On what basis will the best fitting combination of fuzzy sets be chosen?

4.1 Questions 1&2: Metric definition and selection

The system will determine the overall assessment of the time series comparison at the beginning of the content determination task, which will inform the structure of the text realization further along the overall process. This will be determined using key metrics determined from the data behind the two time series, namely correlation and daily variance. Determining the overall assessment upfront allows the system to only display the most relevant information. For example should the correlation be found to be near perfect (e.g. 99.7%) with an average variance of almost nil (0.2%) then the system would only need to communicate these statistics to the user as well as concluding that the fund's performance is almost perfectly tracks the index. However should the correlation come out at 80% (poor tracking) with a low variance (strong tracking) further information would be useful to the user in order to make a decision such as whether there were successive deviations of one time series around another or if there was only one large swing followed by a large correction (necessary to result in an overall low average variation). In this case, the system should then display information about the nature of deviating trends, explaining when, and at what frequency, deviations occurred. The information requirement would be different if the statistics revealed a high correlation and high variance. In this case, one would not expect multiple different trends within the time series but rather a general drift apart of the two time series. It would therefore be more useful to indicate whether the index or the mutual fund's price was higher and provide potential reasons for the drift (such as distributions being declared from the fund).

A key design choice in achieving an accurate assessment is selecting the exact metrics to form part of the assessment. We have already singled out correlation and daily variance as key metrics. Regarding correlation it is reasonable to simply use the overall correlation between the two time series for which the Pearson product-moment approach will be used due to its simplicity. Most complications, which more complex techniques seek to address, are not a concern since no additional measures are computed from the correlation (it is simply the input in a linear fuzzy function).

Daily Deviation (DD) will be defined as the relative indexed price difference between the two time series at any given point according to formula below:

$$DD_{t1} = \left(\frac{Pb_{t1}}{Pf_{t1}} - 1\right) X\ 100$$

Where Pb_{t1} and Pf_{t1} are the benchmark and fund price respectively at t_1 respectively. For the daily variance, a summary statistic needs to be chosen to represent the whole time series comparison. This formula has been written by the author based on widely available percent difference formulae, see (SkilsYouNeed, 2018) for an example.

There are multiple possibilities, such as the average daily, minimum/maximum values as well as potentially the standard deviation of the total DD values. Figure 9 is used to clarify both the problems as well as the reasoning for the final decision. Considering the maximum variance (taken as the largest absolute deviation) as a potential indicator, there are significant advantages. The maximum extent to which the two time series have deviated would be significant to someone analysing the comparison to identify unusual relationships.

> Consider a price comparison with a sudden deviation and then immediate correction (Figure 9B). This would have a minimal impact on the average variance (and correlation) since it did not span many time points however to an auditor this may signal an unusual event and therefore be important. Conversely such a spike may just be an anomaly or inaccuracy in the data and could otherwise lead to a very negative assessment of the comparison. In this case the average variance would take this into account and provide a more stable measure.

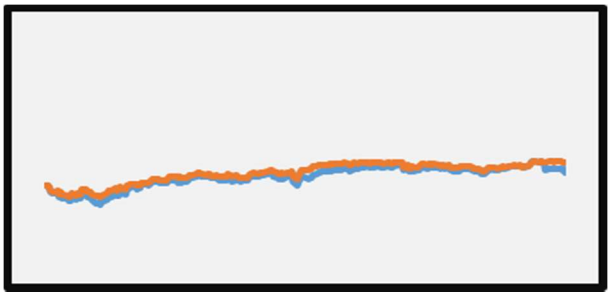
Another possible metric is the standard deviation of the DD. This would provide an indication as to whether a large maximum DD is a once off event or indicative of a large amount of volatility in the DD.

One additional complication is whether the variance exists at the end of the time series. Even if this were the only significant deviation it would be of paramount importance to an auditor since they are attempting to express an opinion over the final balance of the fund. A sudden deviation from the benchmark at year end (i.e. the balance that will be reported by the fund in its audited report), could indicate that the year-end value has been manipulated. At the same time, the ending variance would not provide sufficient information over the entire time series to be used as a determining factor for the assessment.

Note that in graphs A and B, the end DD is equal to Max DD while for C this is not the case. To distinguish comparisons with unusual spikes (which would require focused audit procedures around the time of spike) from a trend of deviations (which indicates a more systemic issue with the comparison), more information is needed. To do this, the average DD is added to the decision criteria which allows the system to distinguish A and C (relatively low average DD) where periodic deviations cause concern from B (high average DD) where there is clearly a systemic diversion of the two prices. Another potential option would be to use the standard deviation DD in place of the max and average DD. For an extreme example such as D (modified from real market data for illustration) however, the standard deviation is not sensitive enough to single spikes to be useful here. It is therefore not considered useful.

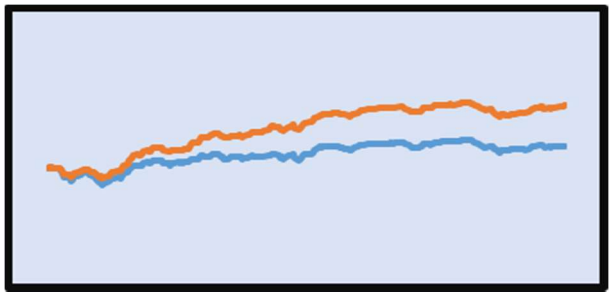
The system will therefore use the below three primary indicators to determine the overall assessment:

1. Correlation
2. Average DD
3. Maximum Absolute DD



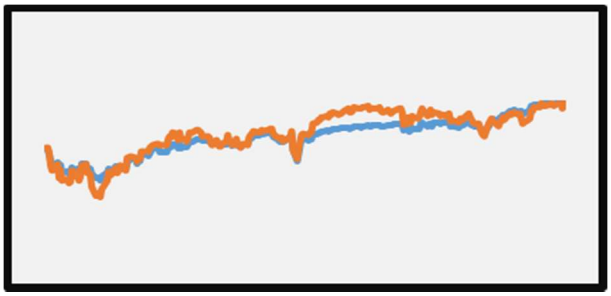
A

Correlation: 98.083%
Max DD: 2.81%
Average DD: 1.07%
Std Dev DD: 0.592%
End DD: 2.81%



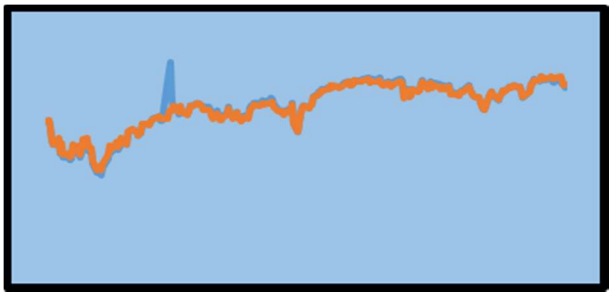
B

Correlation: 98.175%
Max DD: 9.70%
Average DD: 6.16%
Std Dev DD: 2.949%
End DD: 9.70%



C

Correlation: 94.748%
Max DD: 3.63%
Average DD: 0.67%
Std Dev DD: 1.482%
End DD: -0.06%



D

Correlation: 97.914%
Max DD: 8.41%
Average DD: -0.16%
Std Dev DD: 0.905%
End DD: 0.63%

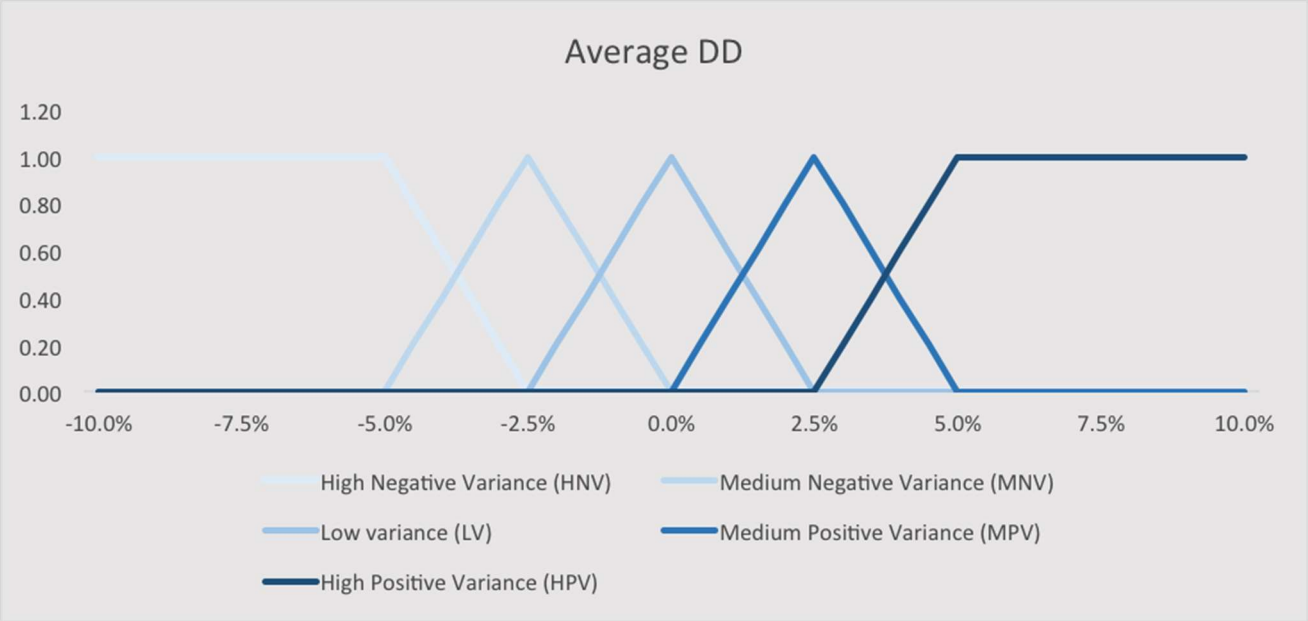
Figure 9: Multiple time series comparisons between Funds and their benchmarks with patterns and potential key summary statistics. A-C are based on real data while D is includes on artificial spike in otherwise real data for illustration.

4.2 Question 3: Fuzzy Functions

Going further, membership scores need to be allocated to each key metric using membership functions. To explain the approach the membership function for Average DD is analysed in detail below with the remaining metric's definitions shown in the appendices.

Average DD has 5 fuzzy sets which can be seen in their graphic representation in Figure 10 along with the boundaries used to define them. Below, the equations for the first two functions (HNV and MNV) are shown for illustration. These are based on the work of (Zadeh, 1965), adapted to this situation:

$$HNV = \begin{cases} 0\ if\ \ X < -100\% \\ 1\ if -100\% \le X \le -5\% \\ \frac{X3 - X}{X3 - X2} if\ -5\% < X \le -2.5\% \\ 0\ if\ X \ge -2.5\% \end{cases} \quad MNV \begin{cases} 0\ if\ \ X < -5\% \\ \frac{X - X2}{X3 - X2}\ if -5\% \le X \le -2.5\% \\ \frac{X4 - X}{X4 - X3} if\ -2.5\% < X \le 0\% \\ 0\ if\ X \ge -2.5\% \end{cases}$$



X0	X1	X2	X3	X4	X5	X6	X7	X8
-101%	-100%	-5%	-2.5%	0%	2.5%	5%	100%	101%

Figure 10: Graphic depiction of Average DD fuzzy sets along with defining boundaries and formulae for High Negative Variance and Mild Negative Variance

The functions are relatively straightforward and consist of multiple linear slopes. Also evident from the depiction is the overlap between functions showing that values of X around the expected range (in this case -5% to 5%) will usually belong to more than one fuzzy set. For example, based on these functions the comparison shown in Figure 9D would have a membership score of 0.94 to LV and 0.06 to MNV.

The percentages used have been determined by the author based on experience in the industry. These values will be assessed when the overall system is tested in order to refine them if necessary.

4.3 Questions 4&5: Decision Calculation

In order to determine the overall assessment, the different fuzzy sets need to be mapped to a final outcome so that they can be directly compared. This mapping is presented in Table 5 and is relatively simple, for each metric the further away the fuzzy set is from a perfect score, the more negative the outcome. In the absence of any further information requirements the results could have been used as

the fuzzy set names for each metric (e.g. replace "HNV" and "HPV" with "Significant Deviations in trend" in Figure 9) however this results in the loss of distinction between HNV and HPV which is utilized further below.

Table 5: The mapping between the individual fuzzy sets for each metric to the overall result

Correlation	Max DD	Average DD	Result
Low	HNV, HPV	HNV, HPV	**Significant Deviations in trend**
Normal	MNV, MPV	MNV, MPV	**Consider further investigation**
High	LV	LV	**Results Satisfactory, recommend no further testing**

In order to arrive at a final conclusion the membership scores for each fuzzy set in each metric are calculated. What remains is to determine how to select the best overall assessment.

> Extending our example from Figure 9D above, the scores for each fuzzy set are calculated, the results of which are shown in Table 6. Arguably the most intuitive method for selecting the final outcome is simply taking the result with the highest average score, as this in theory maximizes the overall fit. If this approach is followed the system will select "Results Satisfactory…" as the result. Considering Figure 9D, this would be an undesirable result as there is clearly a large deviation which would be of interest to an auditor since it represents a possible misstatement. If the result were to be selected based on the highest minimum score (i.e. the decision with the least dissenting score), the system would select "Consider further investigation" as the outcome which would appear to be a more accurate result since, although the comparison overall appears fine, there is an area of interest which should be investigated.

By using the minimum score, the system will try to ensure that every metric has an input into the final result instead of being effectively "outvoted" by two other metrics with extreme scores. This is desirable in unusual situations where one metric (max DD) is able to identify a specific issue (one isolated spike). The minimum score has therefore been selected as the basis on which to determine the result. Where there is a tie in the minimum scores, the next smallest scores will be compared. In the case of a complete tie the more conservative (negative) outcome will be used.

Table 6: The membership scores for each fuzzy set in the example provided by Figure 9D. In addition the average score and minimum score are shown for each conclusion. Highlighted scores indicate the score which would prevail should that metric be used

Result	Correlation	Max DD	Average DD	Average Score	Minimum Score
Significant Deviations in trend	0.00	0.68	0.00	0.23	0.00
Consider further investigation	0.27	0.32	0.06	0.22	**0.06**
Results Satisfactory, recommend no further testing	0.73	0.00	0.94	**0.55**	0.00

4.4 Further consideration: Specific situations

While the above approach does provide a robust, human-centric result, it does not provide any information around any specific exceptions/anomalies in the comparison. Continuing the example of Figure 9D, the system has determined that further investigation is required however no further information is provided. In order to provide this information, an expert matrix is required to provide specific outcomes for particular combinations of metric scores.

> In the example presented in Figure 9D and Table 6 the combination of LNV for the average DD metric, High correlation, and HPV for Max DD indicates the presence of an isolated spike. This is because the HPV in the Max DD metric indicates erratic deviations between the two time series however the LV for average DD indicates that the time series ultimately return to the same position overall. This could just mean that the individual D Ds are moving up and down erratically but cancelling each other out (hence the high Max DD) however the high correlation means there can be very little sustained deviation from the overall shape of each time series. This leads us to conclude that the cause of the HPV was a small number (say less than 3) of isolated spikes. Any more would cause the correlation to drop to normal instead of high based on brief scenario analysis performed by introducing further spikes to the example in Figure 9.

This logic (for this and other situations) is encoded into the specific situations tables which can be found in Appendix A. It is important to note that these specific situations are determined based on the highest score for each metric (e.g. for max DD in Table 6 the selected answer is HPV with a score of 0.68) since the scores assessed are not universal. In the overall result determination, the possible memberships are the same for each metric (e.g. Significant Deviations in trend) whereas in determining the specific situation different memberships are used (e.g. "normal" for correlation but "HPV" for max DD). This is because it is necessary to distinguish between positive and negative variance for Average and Max DD. Since they are not universal, it is not possible to combine them in a fuzzy manner, therefore the highest score is used to determine membership.

An argument could be made to rather create an assessment for each possible combination of the individual scores of each metric and then simply determine the highest scoring fuzzy set in each metric and then to determine the correct combination thereby effectively extending the specific situation identifier to cover all possible outcomes. While this would in theory be simpler (as the two steps above could be combined into one) this would negate the effect of using fuzzy sets. This is because, by simply taking the highest scoring fuzzy set without intermediate processing or combination with other metrics (as is the case in the specific situation step), this would amount to effectively making crisp decisions based on a fuzzy set. Crisp sets could therefore be constructed by taking each low point as the boundary for each crisp set and still have the same result.

> For example in Figure 10 at the x-axis value -3.75% for any x-value lower than that, the HNV would be selected as the answer but as soon as it becomes more than -3.75%, MNV would be chosen (up until -1.25%). The loss of the fuzzy aspect is not desirable in this system as it defeats the aim of trying to implement more human-based computation in the system.

As mentioned in the initial an additional specific situation check was created during initial development to check for the presence of a high ending variance. Figure 11 shows similar examples of both. It was decided to add a specific check since in some cases the overall conclusion may be satisfactory (in which case the system may not show detailed trend information). This however can be

of significant concern to an auditor as previously discussed since ultimately an opinion on the closing balance of the fund needs to be provided.

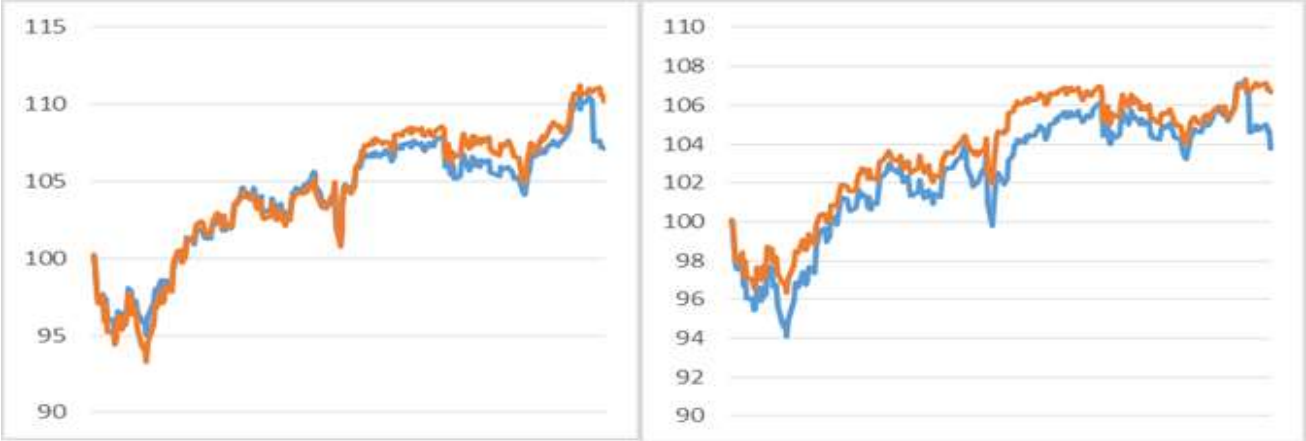


Figure 11: Two similar time series where the maximum deviation occurred towards the end of the period.

A check is performed if the system would otherwise be prepared to state that no significant trends were identified as shown in Figure 12.

1: **procedure** LASTDAYCHECK
2: **if** $dayOfMaxDD \geq NoOfDays[-10]$ and $DDOnLastDay > 0.8 \cdot MaxDD$ **then**
3: $SpecialSituation \leftarrow$ LastDayMaxDD

Figure 12: procedure for checking to see if the max DD occur towards the end of the period, in which case it triggers the relevant special situation

Line 2 checks if the MaxDD value was reached in the last ten days of the period and if the difference on the last day is at least 80% of the MaxDD. The reason for this approach is based on the unusual price movements of the above two examples in Figure 11 . Ten days is used since visually it would appear that this kind of issue could appear within the last ten observations in the time series.

The 80% check is to ensure that in the case where the MaxDD did occur in the last day, that the situation had not been rectified by year end (since this would not achieve the objective of this check). In other words if on day 245 the difference was at its maximum but then corrected by day 251 then there would be no need to report the last day difference.

5 Template Degree

This section examines:

1. The extent to which text needs to be canned rather than loosely defined and subsequently realized through a grammar engine.
2. The information to be conveyed based on the overall assessment as well as the specific situations noted

5.1 Question 1: Canned vs Real NLG

The answer to the first question is largely influenced by the results of Section 4. There are only 3 possible overall results meaning that only 3 different templates are required. This is because the specific situations determined will only require the addition of one or two sentences, and will not affect the core structure of the template. Were these templates to be implemented with a large amount of canned text (with only the specific metrics and numbers being substituted in), then each iteration of a template for the same overall result would lead to almost identical text with only the exact figures varying. This leads to a situation where a user running comparisons on multiple funds may encounter extremely repetitive output across multiple funds which have the same overall outcome. Further, as noted by (Reiter, 2007), a good NLG system should aim to avoid repetition where possible as this can cause a reader to lose attention. It therefore follows that utilizing three templates containing significant canned text will not lead to an effective NLG system. The objective then becomes utilizing as much "real" NLG in each template as possible to introduce more diversity in the texts (although the desirability of repetition will be assessed during evaluation). Another argument for a larger focus on "real" NLG is that ultimately the design should be expandable to other audit procedures, which may inherently require a more flexible approach to text generation. It is therefore beneficial to make the initial design flexible, enhancing future scalability.

5.2 Question 2

5.2.1 Overall Information Structure

In determining how to structure the final template, it is considered optimal to order the information provided by importance. This draws inspiration from (Portet et al., 2009), where only key messages were contained in each paragraph of the summary. The number of paragraphs will not be more than 3 small paragraphs (since there is not as much information to present) and all information will be focused on one message (the comparison of the two time series) instead of multiple events as in the Babytalk system. The focus will therefore be on providing the auditor with the most relevant information first. Following this approach the overall assessment will be presented in the first sentence of the paragraph followed by the most important metric score(s). The next question therefore becomes how to decide on the relative importance of the metrics. The most logical solution is to use the metric whose agreement to the overall assessment is the highest.

> Referring back to Table 6 in the previous section we note that Max DD had the highest score in the chosen category ("Consider Further Investigation" with a score of 0.32). Considering the actual value of each of the metrics in that example (Correlation: 97.91%, Max DD: 8.41%, Average DD: -0.16%), it is intuitive to see that the primary contributing factor to determining that investigation is necessary is the Max DD metric, as the other indicators all point to a satisfactory assessment.

In this situation, (where there is disagreement among the metrics) it is worth highlighting this information to the user as ultimately they are responsible for making the final decision on the comparison. The scores of the metrics which point to another conclusion will therefore be provided. An example text of the above is presented below:

> "*In this comparison further investigation is required before concluding on the valuation of the fund as indicated by a high maximum variation of 8.41% and correlation of 97.91% despite a relatively low average variation of only -0.16%"*

Following the first "impact statement", the summary will go into further detail about the trend comparison. To do this, a summary of the trends identified will be presented. The structure of this information is highly dependent on the information determined in the trend analysis section (refer Section6) and may not be included at all. It is expected that this should take no more than two sentences in order to only provide concise information to the user.

The next component to be included is the summary is a statement around the specific situation identified (refer to Section 4.4). In general the specific situations will have some expected trends attached to them, meaning that once the user has presented with the trend information the specific situation should be easy to accept.

> As an example, if the trend summary reads "there exists a consistent increasing trend across the entire time period with the benchmark outstripping the fund" then the specific situation that follows of "it is highly likely there are distributions from the fund which are not modelled in the comparison" is a reasonable conclusion (refer to Figure 9B in the previous section for the graphic representation of this comparison)

Finally, the system needs to provide any remaining supplemental information that, while perhaps not as important, may still be relevant for decision making by the user. This includes the overall increase in the fund price over the period as well as the key characteristics of the fund, including types of assets invested in (this taken from informal consultation with an industry expert as noted in the initial design section), and finally the indices used to construct the benchmark. Placing this information at the end of the summary allows the user to skip over it easily if they desire.

The information Outline can be structured as shown in Figure 13.

The *[FundName]* Fund *[TrackOutcome]* the Synthetic Index *{[AtAll]}*$_{conclusion}$ *[MetricAgreement]*.

[TrendSummary]
{[SpecialSituationSummary]}$_{SpecialSituationPresent?}$
{[LastDayMaxSummary]}$_{LastDayMax?}$
The overall fund growth was around *[FundGrowth]* over the period. The Fund invests in *[PrimaryInvestments]* {with minor investments in *[SecondaryInvestment]*}$_{MinorInvestments?}$

{Additional Metrics to be considered:}$_{UnmentionedMetrics?}$
{[UnmentionedMetrics]}$_{UnmentionedMetrics?}$

Figure 13: The information skeleton. Regular black text denotes canned words, blue text with "[]" denotes variables (single words or sentences) while text in "{}" indicates text that may not appear (depending on the orange subscript variables, where "?" denotes a Boolean)

5.2.2 Shortened Summaries

As previously discussed there are instances where there is less need to go into significant detail e.g., if the overall assessment is considered "satisfactory". In this case the trend summary would be unnecessary as there would be no notable/significant trends if the system concludes the results as satisfactory. Furthermore the configuration of the system makes it impossible for the system to decide on a satisfactory result and have a finding, since the situation required for a specific situation require at least one metric to have a score that would indicate a failing ("inconsistent") result which would cause the overall assessment to be at best "further investigation". Intuitively, it is sensible that if the system correctly identifies a comparison as satisfactory, less time would be needed to be spent showing information on outliers or problems.

Similarly based on the results of the trend determination it is necessary to state that no significant trends were identified (i.e. if the difference between the two funds is completely erratic) otherwise the summary of trends will become full of too much useless info. This is shown in Figure 14.

```
1: procedure SHORTENED SUMMARY CHECK
2:     if length(TrendSummaryList) > 3 then
3:         TrendSummary ← ShortenedSummary(AllEvents)
```

Figure 14: Pseudocode for the determining of necessity of using a shortened summary of trends.

Line 2: It was ultimately decided that if a trend summary contained more than three separate trends sentences (note that one sentence will usually contain all trends of the same type) that the system will resort to an alternative summary, simply stating the months in which the largest deviations occurred.
Line 3: The method for determining these months is explained under the realization section.

6 Trend identification, classification and aggregation

This section examines:

1. Is there a better approach to identifying trends?
2. What will the membership functions for the trend metrics be and how exactly will they be combined?

After determining the overall assessment and templates for the summary to be communicated, the next step is to determine the trends to be communicated (if any). This step is performed after the template is determined since, in some cases (such as perfect/near perfect key metric scores), communicating such trends would be pointless therefore the template needs to be known before trend identification. Important to remember is that the trends are being plotted based on a time series which represents the difference between the two fund prices. This is done so that this derivative time series can be analysed for trends using the techniques discovered in Section 6.1.

Practically, three steps need to happen before a trend can be communicated. First, trends must be identified from the time series, then classified into categories and finally grouped together. These steps are detailed below.

6.1 Question 1. Trend identification

The proposed method for trend identification noted in the literature research was the use of (Sklansky & Gonzalez, 1980)'s technique for trend identification using linear approximation based on cone graphs. There are many other options for approximation of digital curves (such as time series data) and, since the initial literature did not explore these in detail, they are investigated below in order to determine where this method currently fits in its field and whether there is a more appropriate solution.

(Sklansky & Gonzalez, 1980)'s work set the foundation for a lot of research in the curve approximation field following which additional techniques were developed. (Yin, 2004) provides an overview of these developments and classifies them into three categories (Sequential, split and merge, and dominant point detection), of which (Sklansky & Gonzalez, 1980)'s technique forms part of the first. Each type of technique is considered to have its own strengths and weaknesses. The defining characteristic of (Sklansky & Gonzalez, 1980)'s type of technique (sequential) is its speed and simplicity, at the cost of being dependent on the starting point used.

(Yin, 2004)'s own technique, among others, is a hybrid that approach uses the concept of particle swarm optimization (PSO) to determine the optimum approximation. PSO uses randomly generated particles which each have a fitness score and attempt to follow the particle in the swarm that has the highest score at each iteration, since this should lead it towards the best solution. In another example, (Debled-Rennesson, Tabbone & Wendling, 2004) use the concept of blurred segments which are essentially lines with a certain thickness (loosely its order) which allows them to encompass multiple successive points in a curve. Multiple segments are fitted, each with varying orders and finally, starting with the lowest order and incrementing by one, the optimum fuzzy segment order and corresponding shape is determined.

A common theme that emerged during this research is that the problem which these researchers were attempting to solve was that of approximating an image of some sort, instead of a simple line. This is a more complicated focus area than simply determining trends in a time series as images form more complicated shapes. In contrast, (Dan et al., 2013) focus their research on time series however for the purpose of condensing and summarizing time series data for processing in data mining (one term is used for the entire time series instead of individual trends being identified). The intended result of their research is therefore significantly different from this research problem and cannot therefore be immediately applied or assessed against the current needs of this problem.

What was also clear from the research review is that this field contains far more research than is feasible to review, particularly considering that the scope of this paper is focused on NLG of which this subject matter is not directly a part of. Determining which technique is in fact optimal could be the subject of an entire piece of research and therefore for the sake of simplicity the approach adopted (Kacprzyk, Wilbik & Zadrozny, 2006) based on (Sklansky & Gonzalez, 1980)'s algorithm (shown again in Figure 15) as the former's research aims are very similar to the current problem.

A slight adaption was been made to the above method, specifically the use of a line of best fit. The ultimate result of the above technique is a range of lines inscribed by a cone (as a reminder the figure from the literature research is repeated). One way to determine the trend line is to use the line with the average gradient of the two lines describing the cone. This however has the disadvantage that the resulting line always intersects the starting point of the analysis which may result in a sub-optimal trend line (despite the method correctly detecting those consecutive points as trends). A simple solution to this is to fit a line which best fits all the points under consideration. This is simple to perform through the MatPlotLib package (Hunter, 2007) which provide the gradient and y-intercept of the line which best fits an array of points input. The hybrid approach therefore takes the points which fit the trend determined under the cone method and passes these points to the MatPlotLib package which then returns the line definition. The gradient of this line is therefore used further in determining the classification of the trend.

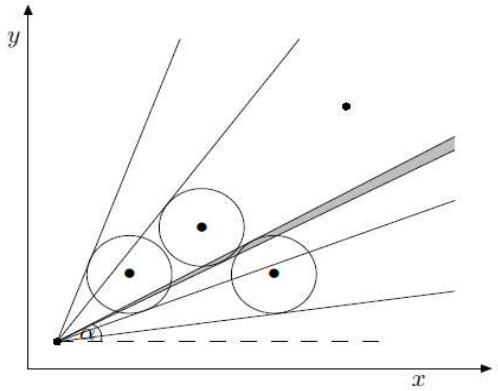


Figure 15: Figure 3 repeated which is an illustration from [Kacprzyk, et al. 2006] showing how cones are built from a starting point which determine a trend at the intersection of subsequent cones

6.2 Additional Consideration: Normalising Time Series and selecting Tolerance

In implementing this trend identification approach, an important decision relates to what tolerance (circle radius) should be used in calculating the trends. The larger the tolerance the larger the resulting cones which then results in more general trends being accepted. As an illustration of this point consider Figure 16 which shows the effect of different tolerances. With the smallest tolerance there is very little generalization and the individual trends don't provide much high level information and make it difficult to aggregate the trends into useable information. Comparatively, using the highest tolerance results in long trends being identified (compare the last third of the time series in each example) however it also loses accuracy as shown by the consecutive downward facing trends in the first third

of the time series, which effectively remove the ability of the system to detect what could be considered a spike.

A complicating factor is fact that each time series comparison will have a different range of values. In order to compensate for this the time series is normalized according to the below formula:

$$Ny_t = \frac{y_t - \min(TS)}{\max(TS) - \min(TS)}$$

Where Ny_t is the normalised value of y at time t and min/max (TS) are the maximum and minimum y values across the time series. The basic normalisation formula (applied here to time series) is widely available, see (Statistics How To, 2018) for an example.

This formula effectively transform all possible values into numbers between 0 and 1. This ensure that all time series comparisons will have a comparable scale on which to base the trend analysis and that the tolerance does not need to be adjusted to each fund comparison. It also has the benefit of allowing the fuzzy functions to be constant.

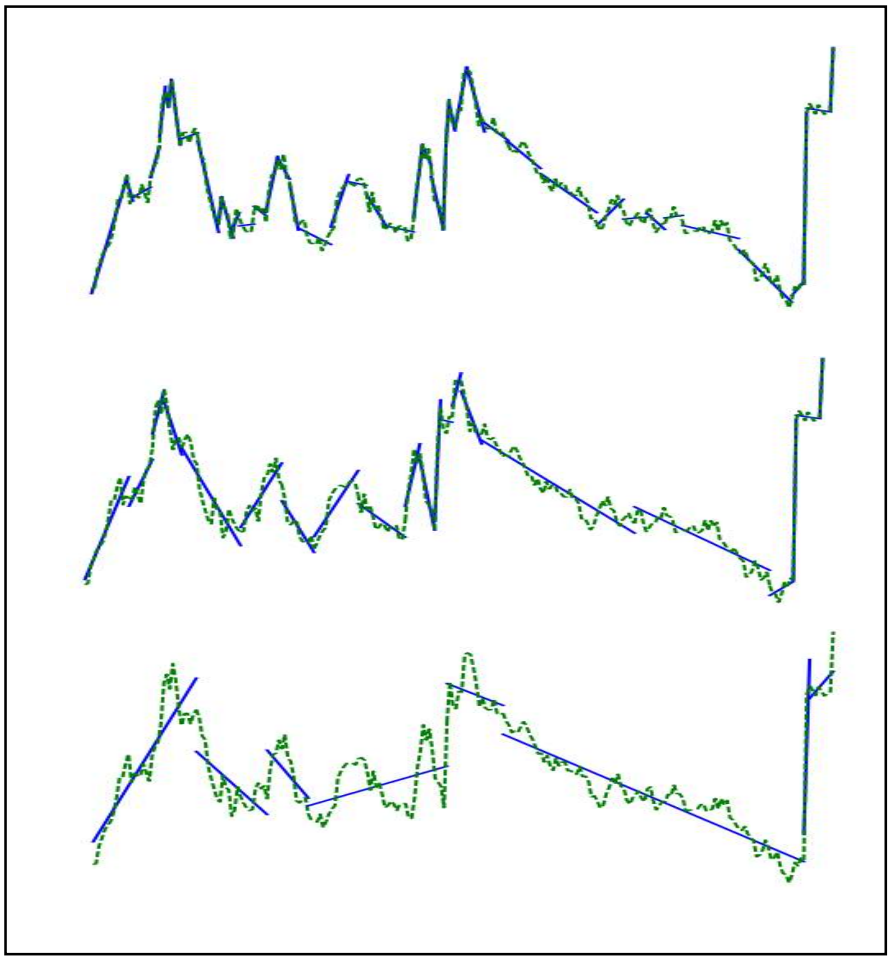


Figure 16: Illustrations of the impact of the choice of tolerance in the trend detection algorithm. From top to bottom the tolerances are 0.05, 0.1 and 0.2. With higher tolerance comes more generalized trends. Note that these graph represent the difference between the time series.

Ultimately, the choice of tolerance is arbitrary and cannot be calculated. Furthermore as far as the research indicates there is no guideline for a tolerance to be used (even with normalised data), and the choice needs to be made by adjusting the tolerance until an acceptable result is obtained. An illustration of the process used can be seen in Figure 16. Successive trend analyses are performed by varying only the tolerance level until the trends identified display the desired trade-off between accuracy and usefulness. This was performed for multiple time series comparison to remove the tendency to tailor the tolerance to one specific example, resulting in poor performance for other time series. The tolerance has been set at 0.1.

6.3 Question 2: Trend Membership Functions and combinations

Once the trends have been identified, they need to be classified in terms of their gradient based on fuzzy membership functions. To do this, membership functions need to be defined which convert a gradient to a variable description of the trend. Similar to the previous membership functions, there are no predefined membership functions that could be found. Indeed, there appears to be research in the field of fuzzy time series (where linguistic variables are plotted against time) such as (Song & Chissom, 1993) however there are no examples of fuzzy functions to describe gradients in time series (even normalised) or other graphs.

Similar to the previous fuzzy functions, the function shown below in Figure 17 has been based off the subjective view of the author with a view to revise them (if needed) once the system is evaluated. While it is difficult to conceptualise, each value on the x-axis is the change in the difference between funds over 1 day expressed as a percentage of the total range of differences across the period.

> For example a 0.01 value on the x-axis means the price difference changed by 1% of the difference between the maximum and minimum differences during the course of 1 day.

More important to note is the clustering of the membership functions. Functions with peaks closer to 0 are more tightly grouped together. This is intentional and represents the nature of gradients in that as gradients increase the marginal angle increase decreases. For example a line with a gradient of 1 has an angle from the x-axis of 45° but a line with gradient 2 (double the first) is roughly 63° which is far

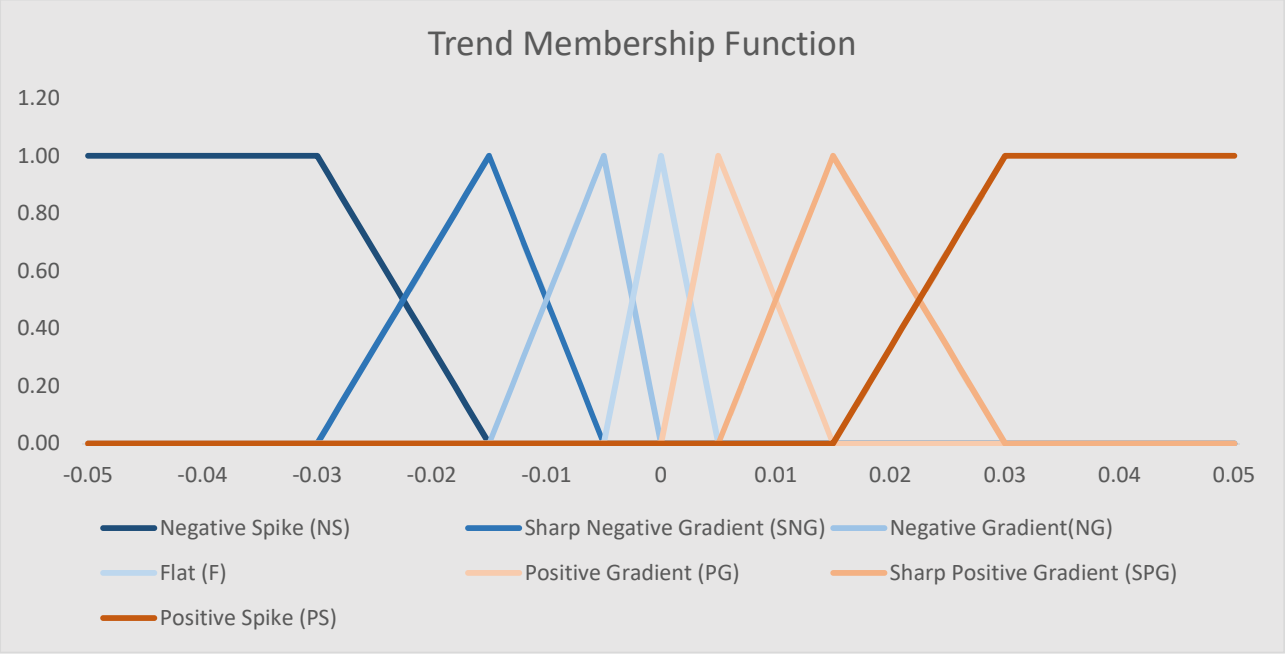


Figure 17: Membership function for trends identified. Memberships to more extreme classification require progressively more extreme gradients.

from double. Larger increases in gradients are therefore required for a visually similar distinction as the gradient increases.

Finally, with the trends established and there is one final piece of processing required in to order to transform the trends into useful information, to classify trends into spikes, jumps and general trends. The procedure for this is shown in Figure 18.

```
1: procedure TREND CLASSIFICATION
2:     for Trend in Trends do
3:         TrendMembershipScores ← TrendFuzzyScores(Trend.Gradient)
4:         if TrendMembershipScores["PS"] > 0.5 then
5:             Trend.IsPositiveJump ← TRUE
6:         if TrendMembershipScores["NS"] > 0.5 then
7:             Trend.IsNegativeJump ← TRUE
8:     i = 1
9:     for index=1 → length(Trends) do
10:        if Trends[index].IsPositiveJump && Trends[index +1].IsNegativeJump
   then
11:            TrendSummary[i] ← spikeup(Trends[index], Trends[Index+1])
12:            Trends[Index+1].IsNegativeJump ← False
13:            i ← i + 1
14:        else if Trends[index].IsNegativeJump && Trends[index +1].IsPositiveJump
   then
15:            TrendSummary[i] ← spikedown(Trends[index], Trends[Index+1])
16:            Trends[Index+1].IsPositiveJump ← False
17:            i ← i + 1
18:        else if Trends[index].IsNegativeJump then
19:            JumpSize ← Trends[index].EndValue − Trends[index].StartValue
20:            if JumpSize > 0.2 · MaxDD then
21:                TrendSummary[i] ← NegativeJump(Trends[index])
22:                i ← i + 1
23:        else if Trends[index].IsPositiveJump then
24:            if JumpSize > 0.2 · MaxDD then
25:                TrendSummary[i] ← PositiveJump(Trends[index])
26:                i ← i + 1
27:        else
28:            GemeralTrendSummary[i] ← GeneralTrend(Trends[index])
```

Figure 18: Pseudocode for the classification of trends with steep gradients into jumps and spikes.

Lines 4&6: Positive and negative jumps are based on a minimum membership function of 0.5 to PS or NS respectively which follows the logic described under Section 4 as belonging more to that fuzzy function than any other.

Lines 10 & 14: A jump in one direction followed by a consecutive opposite jump is classified as a spike. This is classified separately. Whereas a spike indicates that the difference in time series has returned to a normal level, a jump indicates that the change persisted for a period of time and may still persist by the end of the time series which has different implications for an auditor (it is less of an issue if the deviation corrects than if it persists as the former indicates that the problem still persists at the end of the period i.e. reporting period).

Lines 12 & 16: "Used" jumps have their classification removed to ensure that the following trend (index + 1) is not considered as a separate jump as it forms the end of spike but has no opposite jump following it.

Lines 20 & 24: This is done because it was noted that in time series where the difference was relatively erratic many jumps would be identified even though they were not significant compared to the general behaviour of the graph. Setting a minimum change therefore helps to filter out the less significant movements.

For the second step, the aim is to detect other meaningful trends which are not jumps or spikes. This may appear trivial (group all trends with X classification together) however since this information needs to be communicated in a human readable form the method of aggregation is important. For example, if only two trends are identified with the same classification it would be useful to state when each trend occurred however if 10 such trends are identified this becomes undesirable. Instead, trends are aggregated (where possible) sequentially. The pseudocode for this is shown in Figure 19.

```
1:  procedure TREND AGGREGATION
2:      Event ← GeneralTrendSummary[1]
3:      for index=1 → length(GeneralTrendSummary) do
4:          ExistingLen ← length(Event)
5:          NewLen ← length(GeneralTrendSummary[index+1])
6:          ExistingScore ← Event.TrendMembershipScores
7:          NewScore ← GeneralTrendSummary[index+1].TrendMembershipScores
8:          NewEvent.WeightedMembershipScores ← ((ExistingScore ·
    ExistingLen) + (NewScore · NewLen))/(ExistingLen + NewLen)
9:          WeightedScores ← NewEvent.WeightedMembershipScores
10:         for Score in WeightedScores do
11:             if Score > 0.5 then
12:                 Event ← NewEvent
13:             else
14:                 EventList.AddEvent
15:                 Event ← GeneralTrendSummary[index +1]
16:     for Trend in GeneralTrendSummary do
17:         if Trend.LengthInDays < TotalNoOfDays · 0.1 then
18:             TrendSummary.remove(Trend)
```

Figure 19: Pseudocode for trend aggregation. Trends are combined provided they maintain clear membership to one gradient fuzzy function.

This is done by aggregating trends based on their membership functions into trend "events". The system will take the first trend and add the second, computing the weighted average membership function (weighted by duration) for each descriptor (lines 3 to 9). It then assesses the combined fuzzy membership scores to determine if the combined trend still has an inclination towards one classifier (score of 0.5 as for the previous procedure). If this condition is met it adds the next trend. This is done until the combined trend has no clear membership at which point the last trend is excluded and the remaining trends are combined into one "trend event" (lines 11 to 15). As a last step, the combined trends are assessed on total duration to determine if they warrant mention inclusion in the textual summary. 10% of the total period is used initially to determine trends which should be communicated, again based on trial and error.

This is where the power of fuzzy functions is utilized as it allows the mathematical combination of trends.

> As an example if the first trend with duration 25 (days) has membership functions of (PS: 0.4, SPG: 0.6) and the second trend with duration 10 has membership functions (SPG: 0.4, PG: 0.6) then the weighted average membership function becomes (PG: 0.17, SPG: 0.54, PS: 0.29) which allows the combined trend to be considered "Sharp Positive Gradient" where the two composite trends based on predominant membership do not agree with each other.

6.4 Converting derivative information to sensible descriptions

One issue encountered with analysing the trend as one graph (showing the difference between two time series as one time series), is that some information required to provide a meaningful description of the trend is not present in the derivative graph. In particular while it can be determined that the difference between the fund's price slowly became larger than that of the benchmark, it cannot be said if this was due to the fund growing quickly with the benchmark remaining relatively stable or if the fund remained stable while the benchmark dropped. Figure 20 illustrates this problem. In both marked situations the benchmark's price became higher than that of the fund. In situation A however the benchmark grew faster than the fund but in situation B the fund price fell while the benchmark remained consistent. This cannot be communicated simply by analysing the difference in times series, since the behaviour of the two time series cannot be directly observed from the difference of the two. In order to naturally describe this, it is important to be able to compute the difference.

Figure 20 A(top) and B(bottom): Example graph illustrating difficulty in describing changes in times series only in terms of which fund ended up having a larger price. In both situations the benchmark had a higher price that of the fund. In situation A however, the benchmark grew more than that of the fund, while in situation B, the fund dropped while the benchmark remained relatively constant.

In order to get around this issue, the design needs to consider some information from both original time series after determining the trends based on the derivative time series. Additionally, instead of classifying the two times series in terms of which was higher and which was lower the classification needs to be between the "actor" time series (i.e. the one that is deviating the most and causing the difference) from the passive time series (the time series that is relatively more consistent). To do this,

the average gradients of each time series over the period of the trend are calculated and compared. The time series with the highest absolute gradient (i.e. the fund that changed the most) is allocated the role of the actor. The trend is then described in terms of what the actor did according to the following logic:

$$Subject = actor$$

$$Object = passive$$

$$Verb \begin{cases} \text{"grew ahead of"} \ if \ actor = higher \\ \text{"fell behind of"} \ if \ actor = lower \end{cases}$$

Referring back to the example above in both situations the "higher" time series is the benchmark and the fund is the lower. In situation A, the benchmark is the actor while the fund is passive, while situation B is the reverse. Situation A will result in a description of "The benchmark grew ahead of the fund" while situation B will read "The fund fell below the benchmark". As far as the author is aware this approach has not been applied in a similar situation though given its simplicity it is possible this technique is not novel.

7 Realisation

This section addresses the final design question related to realisation:

1. How will the semantic component be structured for further processing into the realisation stage?

7.1 Surface realization

Final realization of the text is handled by SimpleNLG. Functions are defined in java using the SimpleNLG library, which take as input various strings and combine them in a morphologically and syntactically correct way. The total java class is then run as part of a gateway class from the py4j library which waits and listens on a local port for a function name and parameters. This structure is shown in Figure 21.

In general, there is one function per component of the summary. To explain, the overall conclusion function will be used as an example.

The inputs for the function are shown in Table 7 along with an explanation of their meaning. All variables (aside from fundName) are generated in the fuzzy logic stage (refer Section 4).

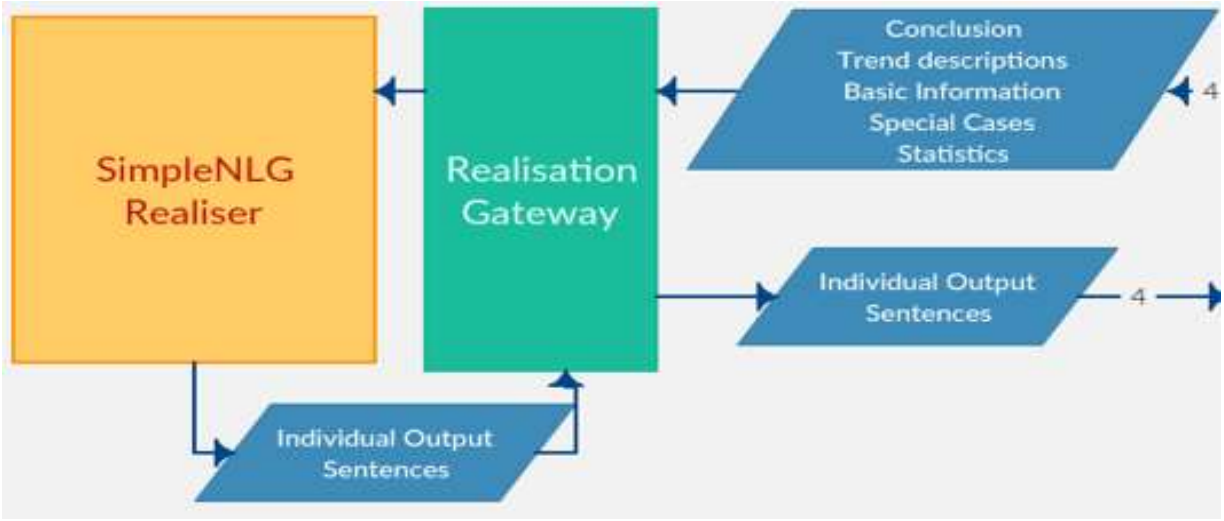


Figure 21: An extract of Figure 7 showing how the primary python design communicates with the Java-Based SimpleNLG module

The skeleton structure of the overall conclusion is shown in Figure 22, with variables being denoted inside "| |". As can be seen there is a combination of canned and variable text following the initial design. The general principle in SimpleNLG is to provide one or more subjects and objects as well as a verb, which will then result in a sentence which can also be joined with others to form more complicated sentences. Switching between "does not track" and "tracks" is performed through switching the negation feature within the SimpleNLG architecture. Similarly the passive feature is used to obtain the "Else" case sentence in Figure 22 which is written in the passive voice.

Table 7: An explanation of the variables used by the overall conclusion function to arrive at a realized output.

Variable Name	Explanation
fundName	Name of the fund being compared
conclusion	the overall conclusion as determined by the system
compromise	Binary parameter to determine if there is disagreement among the metrics i.e. if the overall conclusion is not the class with the highest membership function for all the metrics (not any metrics "first choice"). This can occur when two metrics suggest "satisfactory" and the other one suggests a complete fail. In this case it is likely the system will reach an "investigate further" conclusion although none of the metrics suggested it.
unanimous	Binary parameter to determine if the conclusion reached is that with the highest membership function for all metrics
agree1	If there is neither a compromise nor unanimous decision, this is one of the two metrics which had the highest membership function in the chosen outcome
agree2	If there is neither a compromise nor unanimous decision, this is one of the two metrics which had the highest membership function in the chosen outcome
disagree1	If there is neither a compromise nor unanimous decision, this is the metric which had the highest membership function in a conclusion that was not chosen

For the compromise and unanimous cases, the canned text (e.g. "with differing indications across various metrics") is simply joined to the end of the sentence using "with" as the conjunction.

Some of the variables (compromise and unanimous) could be expressed as Booleans, however for the purpose of communicating between the two languages it was simpler to use strings as these are easier for both programs to understand.

It can be seen that there are a few "if" logic switches within the java function to change the sentence as needed based on what was received from the python function. The other realization functions work in a similar manner with simple IF statements adjusting for different circumstances indicated by the inputs generated at the content determination stage.

The remaining functions and variables are shown in Appendix A: Function/Variable explanation.

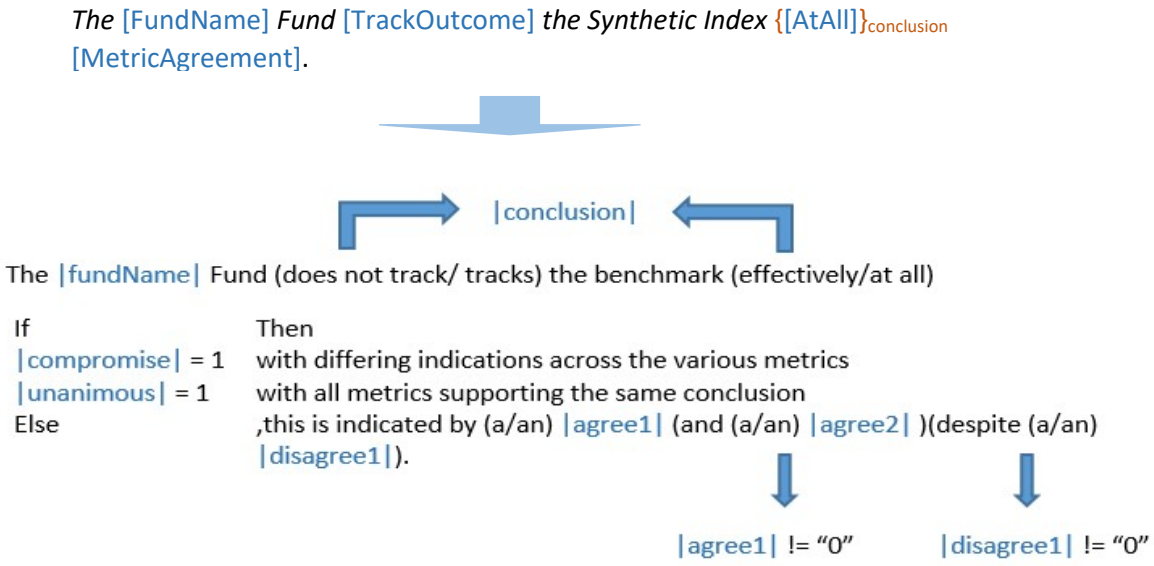


Figure 22: This shows a more detailed view of the original skeleton (top), which can be seen in Figure 13, by providing a decision tree for the conclusion text in the output based on variables previously determined. Words encapsulated by "|" denote variables determined at during runtime. Items in brackets denote variables which can take on one of the options separated by brackets.

7.2 Backend Computations

In order to be able to provide input variables to the surface realizer, numerous calculations and manipulations need to happen on the python side. The notable calculations are explained below.

7.2.1 Time conversion and manipulation

In many cases months are used in the trend descriptions. These need to be obtained only from the day number (e.g., day 153 of 251) in the time series. A complicating factor however is that only working days are included in the time series (financial investments are only priced on working days). The method used is simply to divide the number of days up into 12 parts and treat each as a month. As an example, most funds in this dataset have 251 records which results in around 21 days per month which is in line with the expected number of working days in a month.

In determining the month description, the "early", "mid", "late") approximation is used. The calculation is shown in Figure 23.

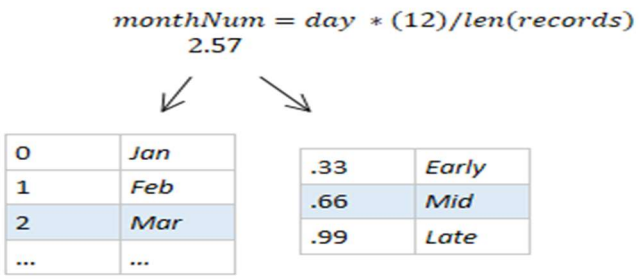


Figure 23: An illustration of the conversion of day numbers into month descriptions. Included is an example.

This method has the advantage of simplicity with the drawback of less accuracy (e.g. it is possible to have the last day of April identified as the first day of May) however considering the final use is to give a general position in the month (e.g. "Early", "Mid", "Late January") not an exact date this lack of accuracy is not considered significant. When looking at a graph of the time series for example being able to tell the difference between late April and Early May (or even Mid April vs Late April) is very difficult.

Additionally, when describing the period of time over which a trend has occurred the number of days need to be converted into a meaningful time measure (days/weeks/months). In order to do this the trend length is classified based on Table 8:

Table 8: The classification of day trend length into more appropriate time descriptors. In the case of 0-5 days, the word "brief" is used to describe the trends, otherwise the description will be based on the number of weeks/months that the trend occurred over.

Type of classification	Time period
"Brief"	0-5 days
Week-based	5-21 days
Month-based	21 + days

Once the number of days fit into one of the next largest descriptor (e.g. 5 working days is a week), the next descriptor is used. This does not extend to years since the comparison period in this case is one year.

7.2.2 Largest Deviation Months

In providing a summary of largest deviations, when there are too many trends (refer to Section 5) the months in which these largest deviations occurred need to be determined. The skeleton for this summary is as follows:

"There were multiple trends and deviations with the largest deviations occurring in <month1>, <month2> and <month3>."

Therefore instead of summarizing all trends, the largest deviations are shown. The challenge however is to determine the largest deviations to display. Only using the largest 3 deviations would often lead to 3 consecutive points (since there is a large chance of the next highest point being sequential to the

highest). In order to resolve this issue only the maximum value for each trend event is compared, the procedure for which is shown in Figure 24.

```
1: procedure ShortenedSummarySelection
2:     for Trends in TrendSummary do
3:         Trend.maxDD ← Absolute(Trend.maxDD)
4:     for i = 1 → 3 do
5:         LargestDD ← TrendWithHighestMaxDD(TrendSummary)
6:         Top3DDs.add(LargestDD)
7:         TrendSummary.remove(LargestDD)
```

Figure 24: The pseudocode for obtaining the trends with the largest three deviations for the period after the trends have been identified in 6.1.

7.2.3 Synonym Choices

Certain domain terms have been chosen by the author based on personal preference and experience in the absence of corpora. Terms such as "track" which in this case is the action of the one time series following another. Other terms could have been "mirrors" or "follows". The logic behind this approach was that this could easily be tested through user evaluation in order to identify choices which are clearly better. This is performed in Section 8.3.4.

7.3 Final design

The design choices made in this section put the final pieces into place for the full design. Data has been converted to information through fuzzy logic, important trends have been identified, aggregated and classified. Finally this information has been converted into text through an appropriate mix of templates and a grammar engine.

8 Evaluation

After presenting the design of the system, it is necessary to assess how it functions in practice. The evaluation of the design below seeks to answer key questions about how the output of this design stands up to expert scrutiny.

8.1 Evaluation Objective

As initially stated, there is a need for a system that can take facts learned from financial information and present it in a format that is easy for someone not trained in data analytics to understand. It is therefore important to assess whether the system is coming to a logical conclusion. At the same time however the output (however accurate) should be understandable to users so that they can utilize the information presented. Lastly in line with general NLG systems there is an expectation that the output should be readable with specific considerations towards fluency and variability in the output text. This leads to the following broad hypothesis:

H1: The generated text is accurate

H2: The generated text is readable and understandable.

H3: The generated natural language summary of fund data makes a positive difference as compared to a graph-only presentation of the processed data, in the context of audits.

H1 and H2 are split since they are determined by separate parts of the architecture (the business and fuzzy logic for accuracy, and the realisation for readability and understandability).

A common contrast made to NLG is a purely graphical approach such as that discussed in [Reiter and Dale. 1997]. H3 is therefore designed around this comparison between graphic only and graphic including text. While it may be possible to design this in terms of a null hypothesis, since the participant base is small (7 people), statistical testing is not considered reliable.

In order to conclude on the above hypotheses the following questions need to be answered

1. Are the conclusions and other information provided by the system accurate? (H1)
2. Is the information provided in the textual summaries understandable? (H2)
3. Is repetitiveness too high or low? (H2)
4. Can improvements be made to word choices to improve the readability of the output?
5. Does the NLG summary either improve or worsen audit impact in terms of: (H3)
    a. Quality
    b. Efficiency

Question 4 does not relate to a hypothesis as no assumption has been presupposed. It is instead a separate evaluation objective.

8.2 NLG Evaluation

Domain experts will be used to test sample output from the system. Since the system has to make fuzzy

decisions, domain experts are needed who can assess whether the decision made by the system is consistent with their expectations. For this purpose having the users test the system itself is neither helpful nor necessary and so the evaluation is performed on the output of the system instead.

Domain experts have been selected from the workplace of the author, the Netherlands branch of one of the "Big Four" accounting firms (PWC, Deloitte, KPMG, Ernst & Young) which are widely known throughout the financial services industry (Accountingverse, 2017) and have a massive client base. Domain experts work in the division of audit known as Investment Management as all the clients (of the audit firm) in the division are involved with managing and growing investments for their clients. These experts have at least 5 years working experience in the industry and, in addition, are qualified chartered accountants (or the country specific equivalent) indicating a high level of technical training and experience in audit. They therefore have sufficient knowledge to evaluate the conclusions reached in the output text.

The participants will be asked various questions via a questionnaire, which will be given electronically (email pdf) via email. Emails will be sent separately, with each participant receiving a participant number to fill in on their questionnaire in order to offer anonymity. The responses are then sent back via email. This study will not require them to be observed and they will therefore be able to complete it at their leisure over a 3 week period. A reminder will be sent to those who have not responded at the end of 3 weeks. Ethics consent was obtained from the Computer Science department prior to conducting the research (approval code: FSREC 59 – 2017).

Prior to conducting a user evaluation there is a preliminary (also broad) hypothesis to test:

Hz: The system outcome is sensitive to changes in the fuzzy functions

Since there is no prior knowledgebase on which to define the membership functions, they had to be created using the author's own understanding of the domain. It is therefore necessary to first conclude on this hypothesis in order to determine whether the fuzzy functions should be specifically evaluated.

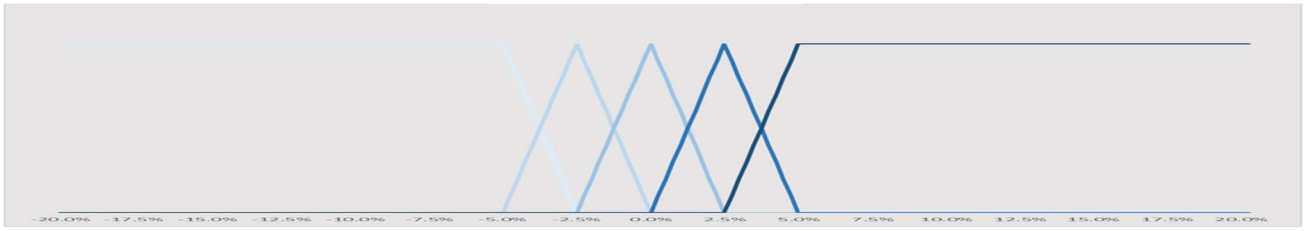
8.2.1 Sensitivity Analysis: Hypothesis Hz

8.2.1.1 Materials and Methods

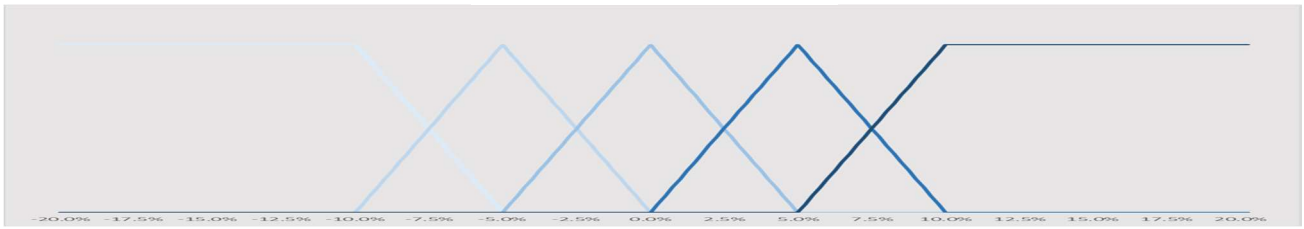
The design of the analysis is as follows, each fuzzy function is adjusted individually while the others are held constant and the impact on the final conclusion is observed. The modification is performed both ways, i.e., by both relaxing the fuzzy function (e.g., where previously a Max DD of 10% would be considered high, it would now be considered closer to medium) and tightening it. This is done by applying a factor of 2 to either double or halve the distance between the perfect value (1 for correlation, zero for both DDs) resulting in a relaxing or tightening of the functions respectively.

Figure 25 graphically illustrates the change between the different versions while Table 9 shows the respective function boundaries. This led to 7 different scenarios; 1 base scenario and 2 scenarios (relax/tighten) for each metric which was run through the designed system and the output recorded and presented in Table 11.

12a: Base Function



12b: Relaxed Function



12c: Tightened Function

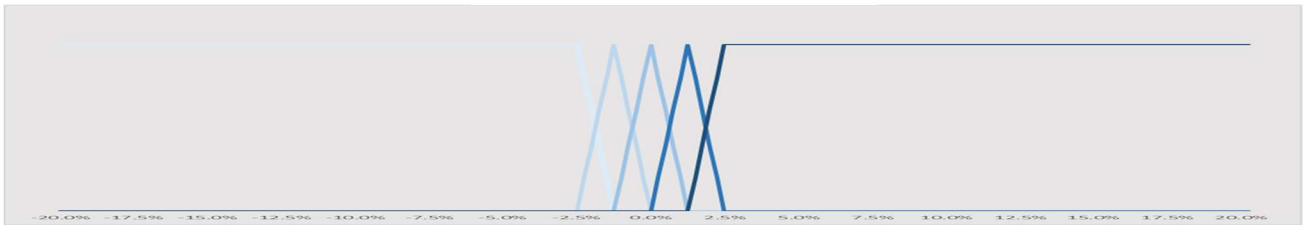


Figure 25: An illustration of the difference between the various forms of the fuzzy functions used in the sensitivity analysis, with the base form first followed by the relaxed and tightened versions.

Table 9: The function boundaries for the sensitivity analysis applied to the Max Daily Deviation metric which correspond to the graphs shown in Figure 25. The base figures are those shown in Table 13.

	X0	X1	X2	X3	X4	X5	X6	X7	X8
Base	-101%	-100%	-10%	-5%	0%	5%	10%	100%	101%
Relaxed	-101%	-100%	-20%	-10%	0%	10%	20%	100%	101%
Tightened	-101%	-100%	-5%	-2.5%	0%	2.5%	5%	100%	101%

Table 10: The function boundaries for the sensitivity analysis applied to the Max Daily Deviation metric which correspond to the graphs shown in Figure 25. The base figures are those shown in Table 12.

Shortened Conclusion	Example conclusion text
Satisfactory	"The Fidelity Growth Company Fund tracks the synthetic index effectively…"
Investigate	"The Fidelity Growth Company Fund does not track the synthetic index effectively…"
Fail	"The Fidelity Growth Company Fund does not track the synthetic index at all."

Table 11: Summary of results from the sensitivity analysis. Blue shaded cell represent changed outcomes in Part A. Finally, the percentage of conclusions changed per metric are shown in part B.

Fund Number	Base	Correlation (Relaxed)	Correlation (Tightened)	Max DD (Relaxed)	Max DD (Tightened)	Average DD (Relaxed)	Average DD (Tightened)
1	investigate	investigate	fail	investigate	investigate	investigate	fail
2	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory
3	satisfactory	satisfactory	investigate	satisfactory	investigate	satisfactory	investigate
4	fail	fail	fail	investigate	fail	investigate	fail
5	satisfactory	satisfactory	satisfactory	satisfactory	investigate	satisfactory	satisfactory
6	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory	satisfactory
7	investigate	investigate	investigate	investigate	investigate	investigate	investigate
8	fail	investigate	fail	investigate	fail	fail	fail
9	fail	investigate	fail	investigate	fail	fail	fail
10	investigate	satisfactory	investigate	investigate	satisfactory	investigate	investigate

Table 12: The percentage of conclusions changed per metric from table 10

	Correlation	Max DD	Average DD	Total
Percentage of outcomes changed	25%	30%	15%	23%

8.2.1.2 Results

The results show that even when altering the fuzzy functions by a multiple of 2 (which is considerable) on average only 23% of the conclusions change. There also does not appear to be any conclusion-dependant sensitivity i.e. the conclusions that changed as a result of the movements are relatively evenly spread out over the 3 possibilities (2 investigate, 2 satisfactory and 3 fail).

8.2.1.3 Discussion of Results

These results are therefore considered indicative that minor changes (less than a factor of two) to the fuzzy functions will not have significant changes to the conclusions.

The implication of these results is that it is not considered necessary to focus significant attention on testing the actual fuzzy functions in the evaluation with users since, while ultimately the conclusion is entirely dependent on the fuzzy functions, making minor tweaks will not lead to significant gains in accuracy. Hz is therefore rejected, resulting in the fuzzy functions not being directly tested in the below evaluation.

8.3 Questionnaire Design Considerations

Questions will asked per graph, with additional questions concerning the output as a whole. The questions to be asked are grouped below in terms of the various criteria noted above, and explained and justified here:

8.3.1 Objective 1: Accuracy

Accuracy of the system is primarily the concern of the content determination stage of this NLG system, as it takes numeric information and transforms it into information using various analytical techniques. (Dale & Mellish, 1998) discuss evaluating such a system, suggesting questions such as "Does it state only information which is true". A binary question such as this is not useful to this evaluation, since the information presented has a degree of subjectivity. For example a conclusion reached cannot be judged as right or wrong since a participant could provide an indication as to how closely the conclusion mirrors their judgement. Additionally, given that detailed data is not available to the user, the description of the trends cannot be assessed as either right or wrong. However a user may note a discrepancy between the starting month in the text summary and in the graphic representation. While this may be a minor difference ("Early May" in the text summary versus "Late April" in the graphic), which could be considered a tolerable and minor inaccuracy, it is also possible that the error could be obvious and serious ("Mid December" versus "Late July"). For this reason a scale will be used with the following options:

1. Highly accurate
2. Fairly accurate
3. Fairly inaccurate
4. Highly inaccurate

A 5-point Likert scale is not used as a participant should not be allowed to provide a neutral answer i.e. they are expected to conclude whether or not the system is accurate/understandable, including the extent to which they consider it accurate/inaccurate. A neutral answer of "the system is neither accurate nor inaccurate" is not logical in this domain.

This question will be posed separately for the conclusion and for the remainder of the information. This is because the conclusion is the part of the system using fuzzy logic which is different from the other components and it is therefore useful to evaluate it separately. If a conclusion is described as inaccurate the user will be asked in a follow up question what their outcome would have been.

8.3.2 Objective 2: Understandability

To assess understandability, and since there are multiple components and sentences within each graph's text, it is useful to ask the participant what portion of the text they do not find understandable. A scale will again be used as follows:

1. I understood all the information presented
2. I understood most of the information presented
3. I did not understand most of the information presented
4. I did not understand any of the information presented

This question will be asked per graph. In addition, in the overall questions stage, the participants will be asked to indicate which parts of the text they did not understand so that they can provide some background information to the answers selected. Understandability can be assessed on an overall level, since similar structure and syntax is used throughout the fund summaries.

8.3.3 Objective 3: Readability

In contrast to the previous criteria readability will be assessed on an overall basis. A key consideration which is to be evaluated under the category of readability is staleness/repetitiveness i.e., How often text is repeated and the effect this has on users. Repetitiveness cannot be tested with only one instance of the text and therefore the texts need to be considered together to determine if there is repetition. To achieve this end users will be asked to rate repetitiveness in terms of the following:

1. Highly repetitive
2. Moderately repetitive
3. Not repetitive

While intuitively it may seem that repetitive text may be undesirable, in this case there is the potential that repetition in certain instances may be desirable. This may be the case when a user wants to quickly scan through sentences to find information they want. In this instance having sentences that change constantly between summaries may not be appreciated by users as it requires them to spend more time searching through the text to find the desired information. To avoid inducing any bias by only asking if they would prefer more variation, participants will be asked if they would prefer more or less repetition. The reasoning behind asking this question will be indicated in order to avoid inducing a negative bias ("repetitiveness" could have trigger a negative bias).

8.3.4 Objective 4: Synonym Choice

A necessary step in realization is choosing the most appropriate synonym to convey the intended message. This choice can be influenced by idiosyncratic disposition of the author, similar to how a normal writer's lexical choices are frequently based on their own idiosyncrasies (Smiley et al., 2016). It is therefore necessary to separately identify the impact of these choices on user's evaluation of the design. While it is not feasible to test the choice of every single word, it is possible to select the more critical words (e.g. key verbs/adjectives) and specifically include those in the assessment.

These words are presented below in Table 13.

Table 13: Words selected for synonym assessment along with their usage in the design as well as possible alternatives

Word	Category	Use	Used in	Possible alternatives
tracks	verb	Used to describe the mirroring of one time series with the other, the better a tracks an index the more alike they are. "follows", "mirrors" or "is similar to".	Overall Conclusion	"follows", "mirrors" or "is similar to"
metrics	noun	Refers to either the correlation, max DD or average DD	Overall Conclusion	"Measure", "Characteristic", "Indicator"
synthetic index	noun	Refers to the single or combination of benchmarks, against which the fund is compared	Overall Conclusion, Trend Summary	"benchmark", "artificial index"

Participants will be presented with one question per word at the overall stage of the assessment asking if they would prefer to see one of the possible synonyms used which will also include the option to specify another synonym. The alternatives are chosen by the author based on experience, since these words are not used in a domain-neutral manner i.e., a thesaurus would yield a large range of synonyms which may not make sense in this context. The option of specifying their (participants) own choice therefore covers the possibility of the author missing a popular alternative.

8.3.5 Objective 5: Testing H3

While the above questions help to provide insight, they are not able to directly test H3. In order to do this participants will be asked to answer the following questions (on an overall basis):

1. Does the textual summary impact audit efficiency compared to the basic summary?
    1. Improves audit efficiency
    2. Make no difference to audit efficiency
    3. Worsens audit efficiency

2. Does the textual summary impact audit quality compared to the basic summary?
    1. Improves audit quality
    2. Make no difference to audit quality
    3. Worsens audit quality

It is the author's experience that a new audit technique is considered useful if it either improves efficiency (i.e. amount of resources expended in performing the technique) or quality (it provides a higher level of assurance than which could previously be attained). The hypothesis could then be accepted if participants consider that either efficiency or quality would be improved by this technique.

8.3.6 Overall feedback

In addition to these targeted questions, the participants will be given an open ended question, asking for suggestions for improvements to the system not already covered. This can be used to make further improvements to the system post evaluation.

8.4 Questionnaire data encoding and metrics

To aid in the evaluation of answers they will be encoded from 4 (indicating high accuracy/understandability) to 1 (Low accuracy /understandability). Each increment of 1 is a different possible answer to question series A, C & E (refer to summary of questions in Table 15 below). The encoding serves to allow numeric analysis.

In order to conclude on the previously mentioned objectives, clear targets are required against which to assess the outcomes. Of the literature reviewed there have not been many instances noted where participants are asked to answer scale-based questions on qualitative aspects such as accuracy and understandability. Studies such as (Smiley et al., 2016) and (Dale, Geldof & Prost, 2003) use comparisons to human text to evaluate their systems while (Castillo-Ortega, Mann & Sánchez, 2011) perform numeric analysis on the output without any external human participants. (Ramos-Soto et al., 2017) do however perform an analysis using a 5 point Likert scale in order to assess the accuracy and relevance of the output of their system. In this evaluation they consider a median score of 4 or 5 to be "good" assuming a low (0 or 1) interquartile range (IQR). A 4 or 5 on the 5 point scale can be considered equivalent to a 3 or 4 on a 4 point scale (since they both represent the two highest possible scores). As such a median score of at least 3 would be considered good for this analysis. In a situation where the IQR is greater than 0 or 1 and therefore indicative of a large spread, the mean will also be considered which should be at least 3 as well (in the absence of other guidance). This applies to evaluation objectives 1 & 2.

Regarding the remaining objectives, these are to be answered with questions which have a maximum of 7 data points as well as smaller scale answers. In these instances, and in the absence of other guidance, a result is considered significant if the number of participants selecting it is at least 2 more than any other option. This however is simply a guideline and a more in-depth analysis may be needed in some cases.

With the above considerations, more specific targets can be defined for the objectives presented earlier and are presented in Table 14.

Table 14: Evaluation questions and the relevant targets needed in order to conclude on research objectives.

Question Number	Description	Metric	Target
1	Are the conclusions provided by the system accurate?	Median (with IQR =<1), else Mean	=>3 (Fairly Accurate)
1	Is the other information provided by the system accurate?	Median (with IQR =<1), else Mean	=>3 (Fairly Accurate)
2	Is the information provided in the textual summaries understandable?	Median (with IQR =<1), else Mean	=>3 (understood most of the information)
3	Is readability and specifically repetitiveness too high or low?	Number of Responses too high or too low	Advantage of => 2
4	Can improvements be made to word choices to improve the readability of the output?	Number of responses in favour of alternatives	Advantage of => 2
5	Does the system improve or worsen audit quality?	Number of responses "improves" or "worsens"	Advantage of => 2
5	Does the system improve or worsen audit efficiency?	Number of responses "improves" or "worsens"	Advantage of => 2

8.5 Evaluation Material

The participants will be provided with an electronic output from the system for 10 different funds as this is considered enough for sufficient coverage. The information for these funds are chosen from real publicly available information with fund prices coming from Yahoo.com, benchmark prices coming from us.spindices.com or yahoo.com (refer to Appendix C for fund names, benchmark indices used as well as their sources). Identifying the appropriate benchmark was done with reference to the fund manager's website (which can be derived from the fund name). In some instances it was not possible to get price data for the correct actual fund benchmark, in which case a similar benchmark was used based on the author's judgement. The exact benchmark is not necessary in this case since imperfections in the benchmark chosen can lead to varying degrees of similarity and therefore provide a more varied range of examples, providing a "stress-test" on the design. Practically, only one modified example was required since the real examples provided good coverage over various deviations and anomalies between the fund and benchmark. This covers the situations which the system could be expected to handle e.g. time series steadily diverging (fund number 4) or time series fluctuating around each other (fund number 8).

In order to test the general hypothesis and give the participants a good frame of reference to compare the output, a purely graphic representation of the data will also be provided along with the three metrics calculated (Max DD, Correlation, Average DD). The participants will ultimately be asked if the system output adds value to the audit evidence in addition to the simple output.

8.6 Summary of Questions

As a point of reference for the evaluation and to provide a clear overview of the question, Table 15 summarises the core of the questions. Multiple choices as well as additional explanations are not included.

Table 15: A summary of all questions in the questionnaire and their question number. Question 2B for example is described on the 2nd data row in table, Question number 2 and question series B.

Question Number	Question Letter	Question
1 - 10	A	How would you rate the accuracy of the conclusion?
1 - 10	B	If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?
1 - 10	C	How would you rate the accuracy of the other information provided?
1 - 10	D	If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate
1 - 10	E	How much of the information did you find understandable?
11		Please indicate which pieces of information you did not understand in the summaries provided
12		Please rate the repetitiveness of the text across the different summaries
13		Would you prefer the same structure of the sentences across the summaries, keep the summaries as is, or more of this sort of variation across the summaries in describing the information?
14		Do you think that the textual summary impacts audit efficiency compared to the basic summary?
15		Do you think that the textual summary impacts audit quality compared to the basic summary?
16		In place of the verb “track” used in the conclusion part of the summary, would you prefer to see:
17		In place of the term “metric” used in the conclusion part of the summary, would you prefer to see:
18		In place of the term “synthetic index” used throughout the summary, would you prefer to see:
19		Do you have any further suggestions for improvement or comments?

8.7 Results

The results of the evaluation are first reported before being discussed in Section 8.8. Following a brief note on response quality, the individual fund question responses are presented first with the general question responses presented afterwards.

8.7.1 Response Quality

Responses were received from all 7 participants. For three participants, a follow up was carried out as inconsistencies were noted in the responses. That is, the participant indicated for a particular fund that the summary was inaccurate for part A of the particular fund question, but when asked for their conclusion in part B, they provided the same answer as that noted by the system summary. Through discussion, it was noted that they had mistaken the example text in the questionnaire as a repeat of the conclusion in the fund summary document ("e.g., The Vanguard Balanced Fund tracks the benchmark effectively": this exact text is repeated in each fund question to clarify where in the fund summary the conclusion is found). A follow up email was sent to all participants explaining this potential misunderstanding. All three participants resubmitted their questionnaires after the clarification and these updated answers form part of the data below.

It was noted that for questions (1C, 2C and 2E) different participants erroneously left the answer to these questions blank. Follow up email correspondence obtained the missing answers, which also form part of the data below.

The question numbers below refer to questions series, an overview of which can be found in table 13 in the previous subsection.

8.7.2 Result Summary

A high level view of the results are provided per evaluation objective in Table 16.

Table 16: The evaluation objectives along with the high level results of the evaluation.

Objective Number	Description	Metric	Target	Actual
1	Are the conclusions provided by the system accurate?	Median (with IQR =<1), else Mean	3 (Fairly Accurate)	Median: 3 IQR: 2 Mean: 2.7
1	Is the other information provided by the system accurate?	Median (with IQR =<1), else Mean	3 (Fairly Accurate)	Median: 3 IQR: 1 Mean: 2.9
2	Is the information provided in the textual summaries understandable?	Median (with IQR =<1), else Mean	3 (understood most of the information)	Median: 3 IQR: 1 Mean: 3.4
3	Is readability and specifically repetitiveness too high or low?	Number of Responses too high or too low	Advantage of 2	More variability: 3 No change: 3 Less variability: 1
4	Can improvements be made to word choices to improve the readability of the output?	Number of responses in favour of alternatives	Advantage of 2	Advantage of 2 for "characteristic" in place of "metric"
5	The system output makes no difference to audit efficiency when compared to the graphical summary	Number of responses "improves", “no difference”, or "worsens"	Advantage of 2	Improves: 6 No Impact: 1
5	The system output makes no difference to audit quality when compared to the graphical summary	Number of responses "improves", “no difference”, or "worsens"	Advantage of 2	Improves: 6 Worsens: 1

8.7.3 Accuracy: Questions Series A & B

Figure 26 shows the results of Question Series A & B. it can be seen that for most (7) funds the participants on average agreed with findings. A mean of 2.7 for example means that the average answer is between fairly accurate (2) and fairly inaccurate (3) but closer to fairly accurate (since 2.5 would be the midpoint and forms the agreement boundary) and falls below the target of 3. Of these 7 two funds (2 & 6) had responses that averaged closer to highly accurate than fairly accurate.

The remaining 3 Questions (4, 9 & 10) had conclusions that the participants on average disagreed with.

Moving to question series B we can see the conclusions suggested by the participants with the size and colour of the circles showing the number of participants who picked a particular conclusion.

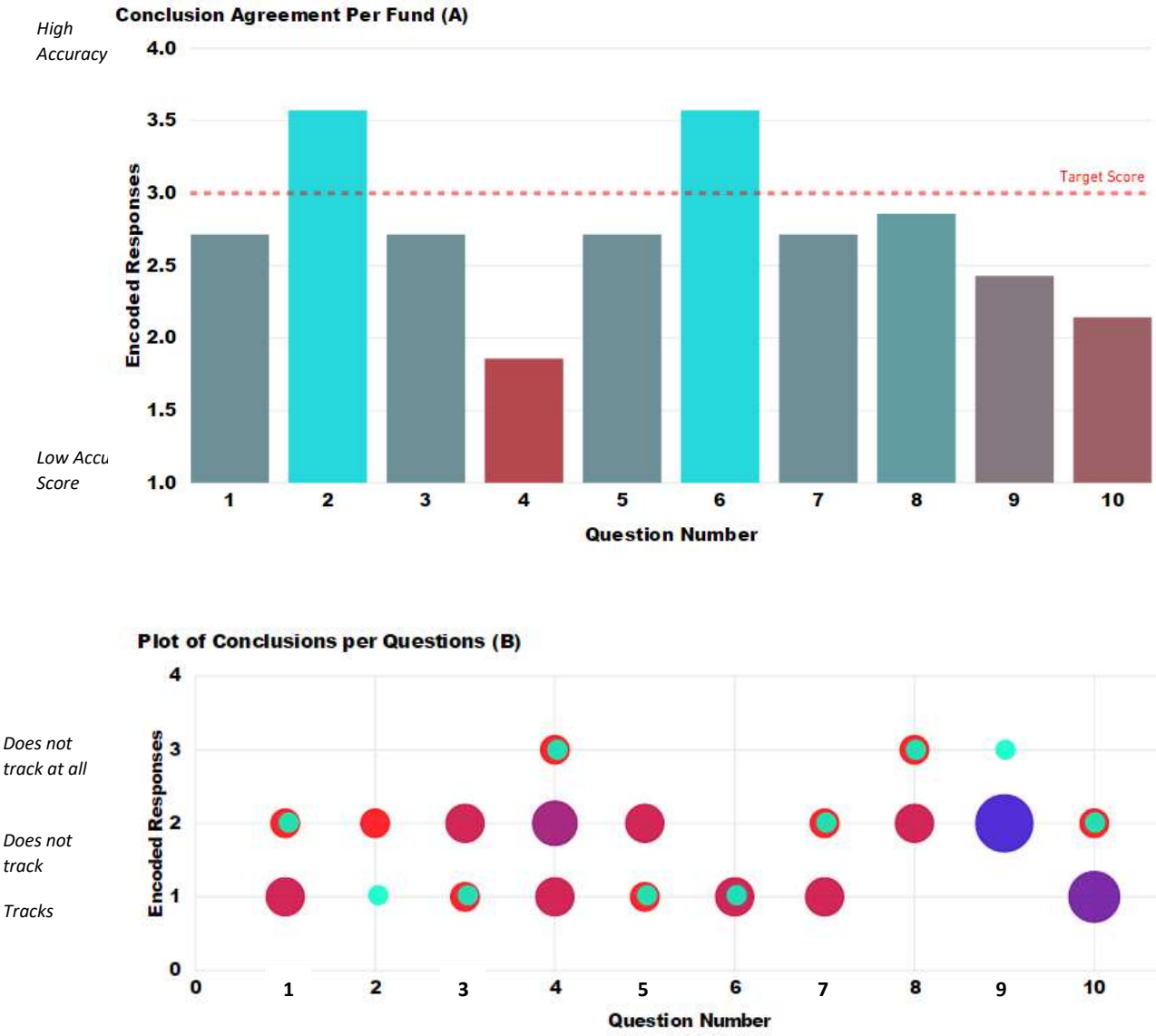


Figure 26: The answers to question series A (top) and B (bottom). Most answers are below the agreement boundary (on average evaluators rate the conclusion more accurate than not) however the majority are also within 0.5 points of the boundary. The summary chosen in 3 out 10 funds is not in agreement with the system. In figure B the green circles show the conclusion indicated by the system.

Figure 26B illustrates some of the findings of Figure 26A. All 7 funds which scored less than 2.5 for part A also had system conclusions that were the same as the majority of participants, as expected (the green circle is in the same place as the largest circle for that fund). Another point to note is that 8 of the 10 funds had only two possible outcomes proposed by the participants, with one of the remaining two funds (Fund 4) having three outcomes and the last (Fund 6) being unanimous in conclusion.

8.7.4 Accuracy: Questions Series C & D

Question series C assessed the accuracy of other information in the summary. Apart from the key statistics above, it can be seen in Figure 27 that for all funds, the participants found the information to be more accurate than not. Only four of the funds met the target. The funds with the highest other information accuracy are also the funds which had the most accurate conclusions according to the participants. The same cannot be said for the funds with the least accurate other information scores, with only Fund 4 being consistently inaccurate according to participants (funds 9 and 10 are in the top 5 of most accurate other information).

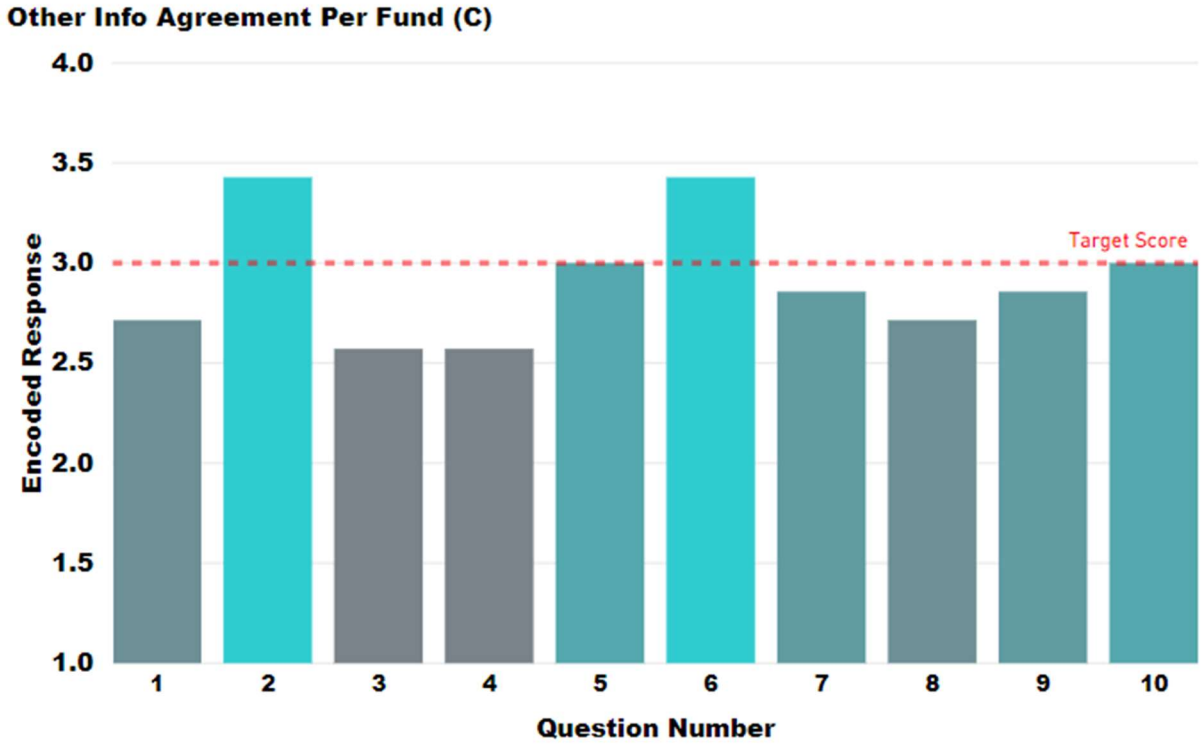


Figure 27: The answers to question series C. All fund summaries were rated closer to accurate than not but only 4 met the target score

Question series D provides qualitative reasons for why participants rated the other information as inaccurate. Going through the lowest rated funds, the participants comments are summarised in Table 17. Aside from two responses, the three same participants provided all the inaccurate ratings above.

Table 17: A summary of responses provided by participants when indicating that the other information provided by the summary was not accurate.

Fund 3	Of the 3 participants who rated this summary "fairly inaccurate", two cited the failure of the summary to mention the large deviation that occurred at year end as the cause (participant 6: "*Early December large deviation not mentioned?*"). The other indicated that they were not able to understand what was meant by some of the daily deviation metrics (participant 5: "*Don't understand the remark on positive maximum daily deviation*")
Fund 4	One participant rated the summary highly inaccurate as they disagreed about the possibility of distributions being declared from the fund (participant 6: "*How would you come to the indication of distributions (likelihood)? Should this not be visible in a spike?"*). Another participant rated it fairly inaccurate as they do not agree with the use of vague terminology such as "*high/very high*"
Fund 1	Of the 3 participants who rated inaccurate (fairly inaccurate) to this summary, two cited the failure of the summary to mention the large deviation that occurred at year end as the cause. The last cited the lack of reasons for the deviations as the problem (participant 7: "*There is no further description of the reasons for the average deviation/ maximum deviation. In addition there is no description why the benchmark is more in line at the end of the year.").*
Fund 8	One participant rated the summary highly inaccurate as they disagreed about the phrase "complete lack of similarity", (participant 7*: "I*n my opinion the statement 'complete lack of similarity' is not in accordance with my expectation."*). Another rated it fairly inaccurate as they believed the largest deviation was not mentioned.

8.7.5 Understandability: Question series E and Question 11

The last of the individual fund questions shows the participants rating of the understandability of the texts, the results of which are shown in Figure 28. Following a similar approach to the accuracy presentation, all of the summaries were considered understandable by participants for at least most of the information. It should be noted that the summary for Fund 4 is again one of the worst scored (it was also within the top 3 least accurate conclusions and least accurate other information).

Relating to question 11 "*Please indicate which pieces of information you did not understand in the summaries provided",* only three participants provided responses which are as follows:

- *"How to read additional metrics? What is link to conclusion that fund tracks effectively, ineffective or not at all? Is there a 'norm'?"*
- *"Information provided is fairly accurate but not always complete. More information on i.e. automated analysis of deviation could be given."*
- *"On what basis is the correlation calculated? On what basis is the maximum and average deviation calculated"*

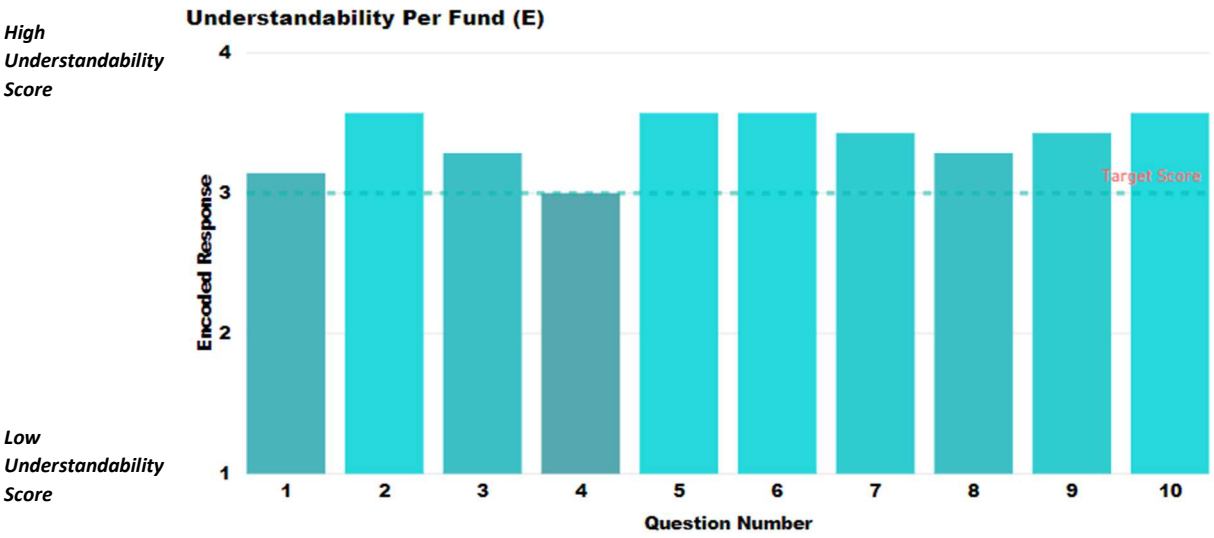


Figure 28: The answers to question series E. On average most of the information presented for all summaries was understandable.

8.7.6 Repetitiveness: Question 12 & 13

Moving to the question of repetition, the answers are presented below in Figure 29:

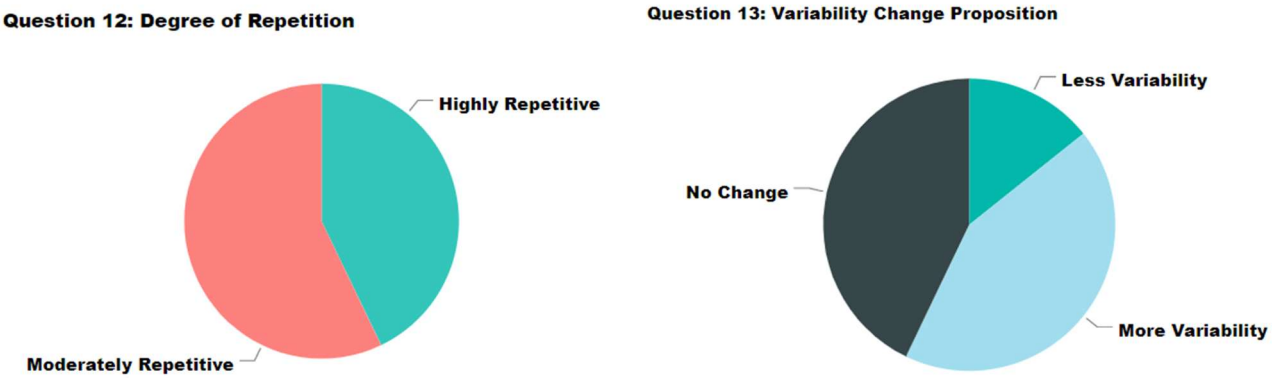


Figure 29: The answers to questions 12 & 13. While all participants considered the text at least moderately repetitive there is an almost even split as to whether more variability is desirable.

Regarding the perception of repetitiveness of data, the participants consider the data at least moderately repetitive. When asked if they would prefer more, less or no change to variability however there was a split in opinion between more variability being necessary and not. It should be noted that there is no clear pattern to the participants' answer, i.e., of the three participants who answered "highly repetitive" to question 12 each of them gave a different suggestion for a change to the variability in question 13 i.e., two did not want to reduce the repetitiveness nonetheless.

8.7.7 Audit Impact: Question 14 & 15

Figure 30 shows the answers to questions 14 & 15 and shows that all participants except one consider the textual summaries to have a positive impact on audit efficiency and quality. Participant 4 indicated that no impact is made on efficiency and that quality is worsened.

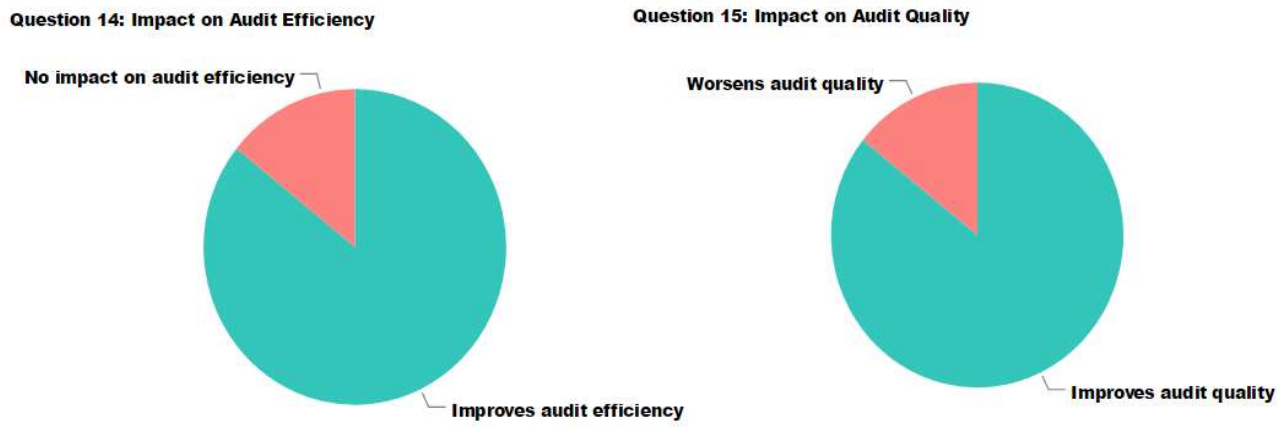


Figure 30: The answers to questions 14 & 15. All but 1 participant indicated that the textual summaries had a positive impact on audit efficiency and quality.

8.7.8 Alternative Word Choices: Questions 16 – 18

The final categorical question concerned alternatives to the terms "track", "Metric" and "Synthetic Index", the results for which are shown in Figure 31.

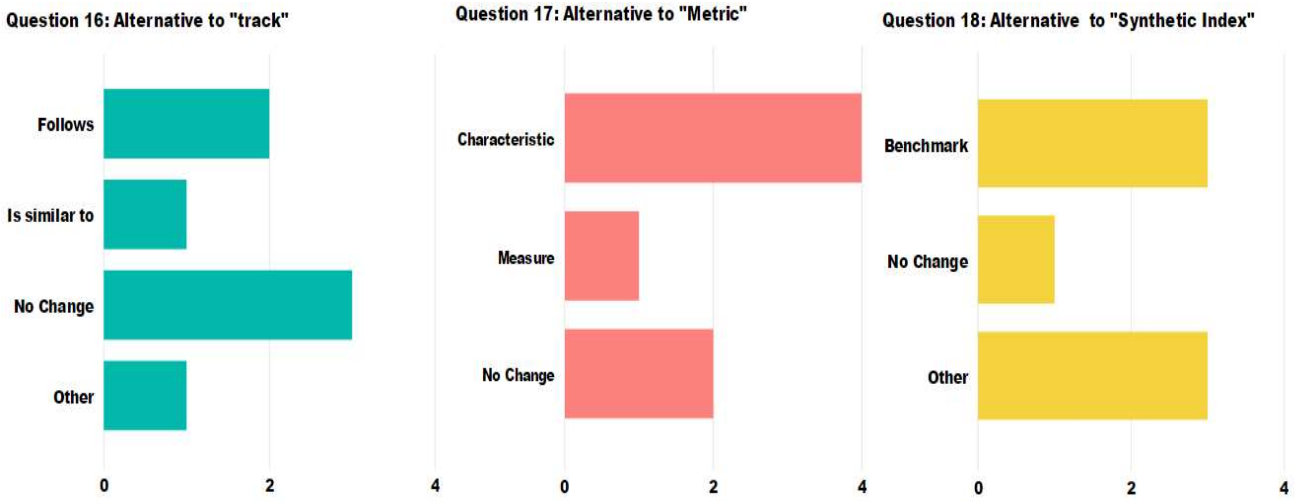


Figure 31: The answers to questions 16, 17 & 18. Most participants preferred the word "track" to any other single alternative but had preferences for other terms in place of "Metric" and "Synthetic Index".

While most participants had alternative preferred words to the existing terms, for the term "track" there were more in favour of not changing than there were for any single alternative. For Metric however "Characteristic" was preferred by the majority of participants and for "Synthetic Index" there was a

split between Benchmark and other suggested alternative. The three participants suggesting other terms proposed the following terms:

- Synthetic Benchmark (suggested by two participants)
- By the fund indicated benchmark (e.g. from the prospectus)

8.7.9 Other Suggestions: Question 19

Finally participants were asked for other suggestions they would like to see to improve the system output. The answers provided are shown in Table 18 below.

Table 18: The other suggestions for improvement provided by the participants.

Participant Number	Comment
1	*"In the textual summaries I would like to see more clearly which value deviation requires audit follow up"*
2	*"Instead of using the word "effectively" in the explanations, I'd suggest to use the following:* *The evolution of the Fund can be explained by the benchmark due to its high correlation(for example)"*
3	none
4	*"Information provided should be similar (setup) for every summary (template). It would be helpful if the system could provide information on outliers derived from observable data."*
5	*"add benchmark growth figures in addition to fund growth figures"*
6	*"I would like to see that follow up procedures are indicated.* *e.g. The engagement team needs to inquire the large spike/deviation early December (Vanguard Wellington)"*
7	*"Question 10; does not track is quite conservative,* *there is only one item which is not in line with the synthetic benchmark. Therefore make another sentence"*

As background to participant 7's response (the participant discussed this point before submitting), they believe that there should be some middle ground between does not track effectively and tracks effectively for this particular (modified) fund.

8.8 Results Discussion

Following the presentation of results, they are discussed below with an interpretation of their meaning. A reflection of the evaluation method is provided before the results are analysed per evaluation question.

8.8.1 Objective 5: Audit Impact

The impact on audit efficiency and quality was almost unanimously considered to be positive, with the responses indicating an improvement being well in excess of the advantage of 2 determined above. While the number of participants makes it difficult to perform a statistical analysis of this result, it is definitely encouraging for the design as a whole. An interesting point regarding participant 4 who provided the negative rating (for both quality and efficiency), is that they provided conclusion accuracy and understandability ratings which were more favourable than the average rating across the participants (although their scores for the other information accuracy were less favourable than the average). In addition, it is not clear from their suggestions for improvement what the cause for considering the impact unfavourable is. Follow up was sought with the participant however contact could not be made. Other than that however no further action is considered necessary in this regard.

8.8.2 Objective 1: Accuracy

8.8.2.1 Conclusion Accuracy

The target for the median score was met with a score of 3 however since the IQR is higher than 1 and the mean score is below 3, the conclusion for this objective is that the accuracy of the conclusions provided is not sufficient. What remains to be discussed is whether the information from question series B can assist in refining the accuracy of the design.

In order to effectively interpret question series B, an analysis was performed whereby additional responses to series B were interpolated from answer A (since an answer is only required if the participant disagreed with the conclusion). The algorithm to do this is as follows:

1. Is answer to XXB blank and is the answer to XXA 1 or 2?
2. If so replace answer to XXB with the conclusion suggested by the system.
3. Else leave answer as is

The purpose of this interpolation is to gain a more complete picture of which conclusion each respondent considered the most accurate.

> As an example, for Question 2b if only the respondent scores are considers there are 2 “votes” for “does not track” and none for “tracks”. This ignores the other 5 participants who indicated that the system’s outcome of “tracks” is at least “fairly accurate” and would implicitly also choose “tracks” if required to answer. Only when these are combined is it possible to see that in fact there are 5 participants implicitly voting “tracks” versus two voting “does not track”. This allows to conclude the most popular conclusion is in fact “tracks”.

The underlying assumption behind this inference is that if a participant answers "fairly accurate" or "highly accurate" to series A, then they arrived at the same conclusion from the data as the system did. The result is shown in Figure 33 below where it is used for further investigation into the benefits/costs of making changes to the system logic. In hindsight the better solution would have been to make Question series C mandatory.

With a clear view of the "correct" conclusion according to the participants for all funds (with the exception of Fund 4), a target is available which the updated design should achieve. For two (Funds 4 & 9) of the three funds which had an average inaccurate score, the conclusion was reached by the system through the use of the tie-breaker algorithm described in section 4.3. The tie-breaker algorithm, in the absence of a highest minimum score for one conclusion, takes the next lowest score for each conclusion and looks for the highest. As it is responsible for these "poor decisions" instead of the primary algorithm, it may be possible to improve the accuracy of the conclusions by making an adjustment to this tie-breaker algorithm.

One possible adjustment would be to default to the middle conclusion ("does not track effectively") in the case of a tie. The reasoning behind this is shown in Figure 32. In all cases the conclusion "Does not track effectively" is either one of the outcomes suggested by at least one metric, the middle ground in the case of polar opposites (third line) or each metric suggesting a difference conclusion (fourth line). This logic seems to be supported by the data from the participants. In both instances of this tie-breaker participants went for the middle conclusion.

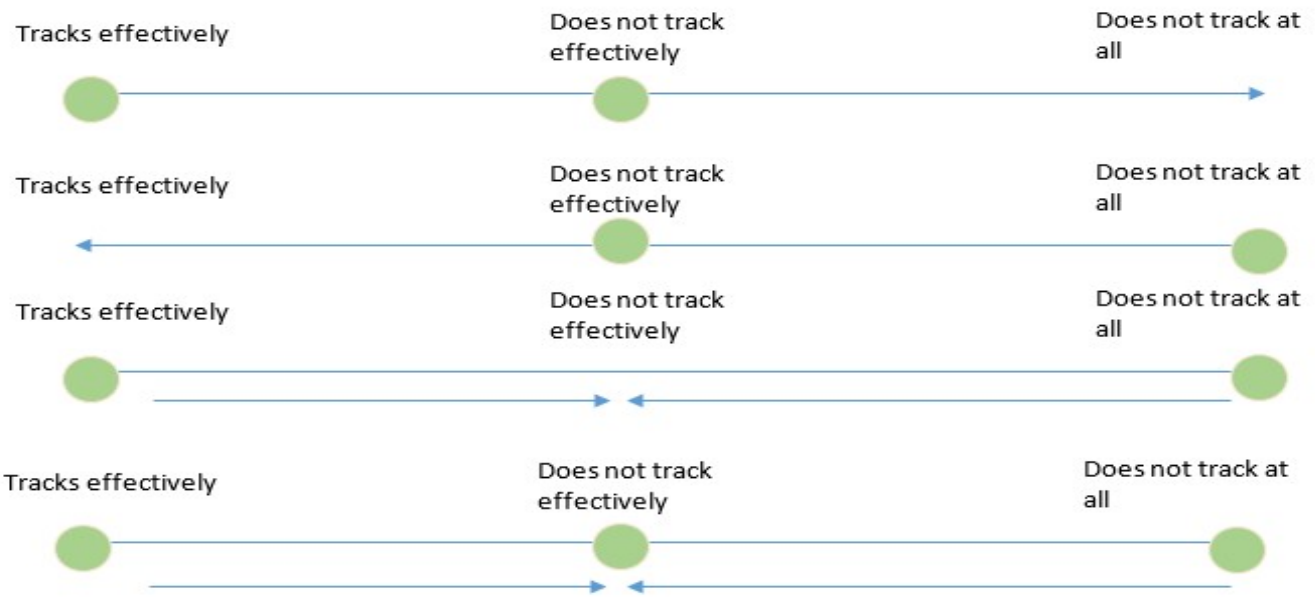


Figure 32: Possible outcomes where a tie-breaker could occur.

While this may be a simple solution to the problem there is the possibility the impact on the other (more accurate summaries) has not been assessed. Aside from funds 4 & 9, Fund 8 also made use of a tie-breaker to reach its conclusion, which was the most negative one. This conclusion was the most accurate according to participants. Making the proposed change to the tie-breaker algorithm would therefore result in a middle outcome for this fund which would result in a loss of overall accuracy. This is illustrated in Figure 33 below which shows the updated system conclusions in yellow circles.

Note that the proposed change in conclusions would result in the accuracy gain (moving from red to purple circle) in Fund 9 being almost exactly offset by the loss in of accuracy in Fund 8 (purple to red) with the marginal improvement only being the increased accuracy in Fund 4 (which is marginal due to the conflicted opinions of the participants).

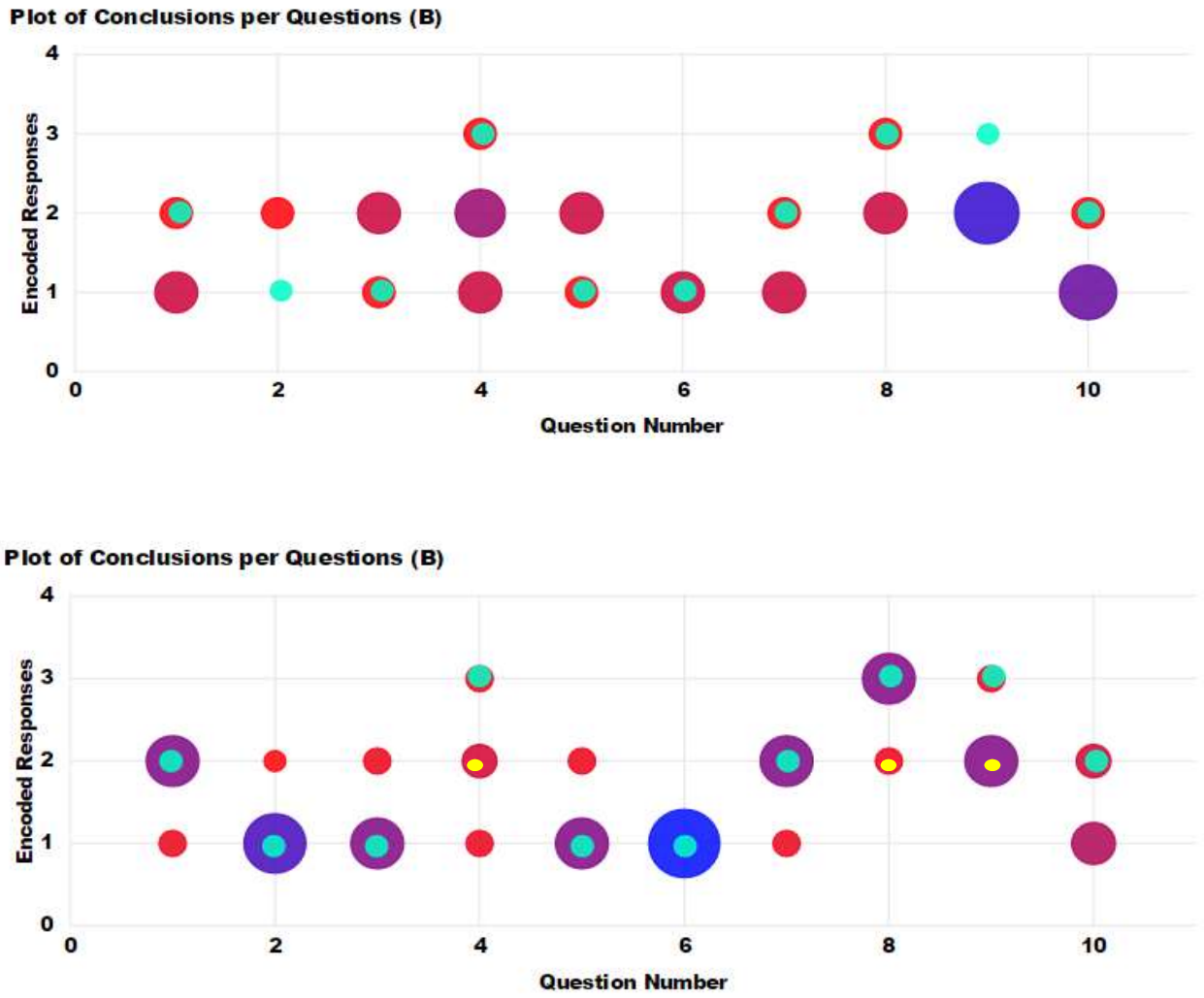


Figure 33: The original 26B (no analysis) is repeated above with the updated conclusion based on the interpolation below. The below figure also include the adjusted system outcome if the change to the tie-breaker were implemented

While an argument could be made that any improvement in accuracy is desirable in this case, given the small number of observations with which to work (2 examples supporting the change and one against) there is a significant risk over-refining the system (similar to overfitting a model to one set of data in statistics) since we are considering only a small part of a sample which may not sufficiently represent the whole population to support this change. For example, if there had been just one other fund like Fund 8 in the test then the result would have indicated an overall accuracy loss from making the proposed change. Given this level of uncertainty, it does not seem advantageous to implement this change.

It was then considered if there was another possible algorithm that could get to the same conclusion as the participants for all the funds (including Fund 8). A simple comparison between funds 8 & 9

performed in Table 19 however shows that while the participants chose different conclusions for the two different funds, based purely on the three metrics' membership functions, they are almost identical.

This comparison of funds 8 & 9 also raises the question as to whether these three metrics or the setup as a whole is complete enough to provide an accurate solution. Based on these results it seems that additional input is required to improve the accuracy of the design.

Table 19: The membership scores for Fund 8 & 9. Although the participants chose different answers the membership scores are almost identical

	Fund 8			**Fund 9**		
	Correlation	Max Deviation	Average Deviation	Correlation	Max Deviation	Average Deviation
Does not Track at all	1.00	1.00	0.00	1.00	1.00	0.00
Does not track effectively	0.00	0.00	0.98	0.00	0.00	0.87
Tracks effectively	0.00	0.00	0.02	0.00	0.00	0.13

8.8.2.2 Other Information Accuracy

Moving to the other information provided by the summary, there were less accuracy concerns here than compared to the conclusion, however on average the accuracy was still assessed to be between fairly accurate and fairly inaccurate. In terms of medians, the same score of 3 is obtained with a narrower IQR of 1 this would mean that the requirement for this evaluation objective is met and we can conclude that the other information is sufficiently accurate. This does not mean however that it cannot be further improved.

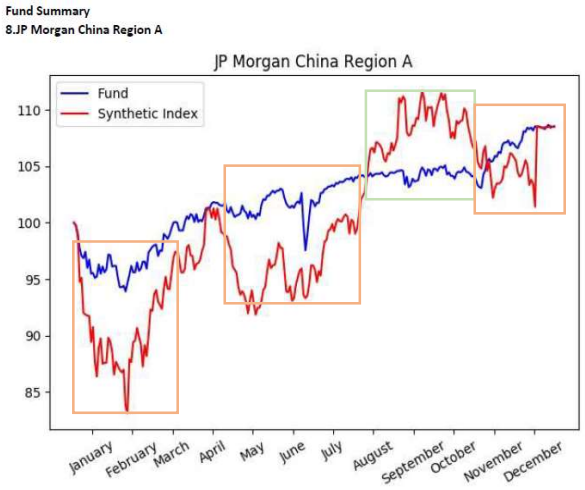


Figure 34: The graphical illustration of Fund 8 including the largest deviations mentioned by the textual summary (orange box) as well as another deviation not mentioned (green box)

Examining the qualitative feedback, the issue of not mentioning deviations which the participants considered significant was noted (funds 1, 3 & 8 in Table 17) was the common theme. This has been caused by the use of 3 as the default number of deviations to describe when the maximum number of describable trends has been reached (refer section 5.2.2). A good example of this can be seen in Figure 34, which shows the graphic output for Fund 8 as well as which deviations were mentioned in the textual summary and which were not. The participants noticed and evaluated the

deviation between August and October as noteworthy, however the system did not since it contained a maximum daily deviation that was not as large as the other three. Had the system been setup to display 4 deviations it is highly likely that this would have been included. However for another fund where only 3 significant deviations occurred, the summary would include one extra deviation which is not significant. There is therefore a need to develop a technique for determining the number of trends to mention based on the nature of the deviations in each fund.

The other comments mentioned by the participants are unique to that participant and are not shared by the others, and therefore no further changes are considered necessary. One comment however (not understanding how the metrics are calculated) could simply be handled through the inclusion of basic tool documentation (including methodology for calculating the metrics) which would form part of the audit documentation.

8.8.3 Objective 2: Understandability

Regarding understandability, the feedback from participants placed the average understandability rating between "I understood all the information" and "I understood most of the information". The median of 3 with low IQR as well as mean of above 3 indicate that the system's understandability is sufficient. In addition, only three participants provided details of information they did not understand (one of which was already mentioned under the preceding paragraph). There is however a comment from participant 6 who wanted a clearer link between the "additional metrics" and the conclusion. Email follow up with the participant clarified that they wanted to know how the system had gotten to the conclusion based on the additional metrics. This would be resolved through a "theory paper" on the software i.e., audit documentation that explains how the audit technique works in a less technical manner. The last comment called for an automatic analysis of the deviation which is taken to mean that they would like reasons/insight to be given into the deviation.

8.8.4 Objective 3: Readability

While participants understood almost all of the output and did not have many suggestions, when asked about repetition many had alternative wordings as well as a generally divided opinion about whether more variability would be desirable.

While the participants agreed that there was at least some degree of repetition (expected in almost any NLG application), they were divided on whether more variability would be desirable (an even split), and the necessary advantage of 2 was not obtained. Based on this result it would not seem wise to make a change to the variability (recall there is an expected benefit to similarity in this application). Note that there is a potentially inconsistent answer provided by Participant 6 as they indicated that they considered the text "highly repetitive" but then suggested less variability in the following question. It is therefore possible they intended to answer "more variability" for question 13. Email follow up was performed where the participant was asked for clarification. As a result the participant indicated that they stood by their answer to question 13 ("less variability") they indicated that they would like change their answer to Question 12 to ("moderately repetitive"). This change has a minimal impact on the

analysis of question 12, shifting the general consensus further towards “moderately repetitive”. This does not change the conclusion that all participants consider the text at least moderately repetitive.

8.8.5 Objective 4: Synonym Choices

Moving the alternative word choices, we see that the participants are in favour of maintaining the term “track” over any one other alternative. For the other two terms (“Metric” and “Synthetic Index”) there is a fairly clear preference for “Characteristic” instead of “Metric” (an advantage of 2) while the situation for “Synthetic index” is not as clear. There was a split between the provided alternative “Benchmark” and personal suggestions. It is important to note however, that all the suggestions included the term “benchmark” somewhere in the suggestion. There is therefore overwhelming consensus (6 to 1) that some variation of “benchmark” is preferable to “Synthetic Index”, making it the superior choice for the design.

8.8.6 Other Suggestions

Discussing the responses to question 19 not already addressed above, an interesting suggestion was to include specific deviations and issues for the audit team to look into. This would effectively form a checklist for the audit team to complete before finally passing judgement on a fund. This would be likely be a useful change to implement.

Another suggestion requested the benchmark growth figures to be shown in addition to that of the fund. While the use of this information does not appear particularly significant, it would not take up much additional space in the summary (e.g. “The overall fund growth was around 8.54% over the period **compared to that of the benchmark of XX%)** where red text indicates added content and could be seen as a low benefit, low cost change to implement.

Another comment suggested a complete change of the conclusion structure as well as one asking for an additional conclusion type are not considered desirable to implement as this would be a significantly larger change with only one participant behind the suggestion.

8.8.7 Summary and hypotheses

The conclusions can be summarised based on evaluation objectives are summarised in Table 20. While the results are largely positive, there is a still a concern relating to the accuracy of the conclusions generated by the fund. This is a large concern, as providing accurate conclusions is one of the key functions of this designs. The evaluation results however could not identify a clear solution for resolving the lack of accuracy however the application of larger scale evaluations as well as the application of machine learning in future could allow for further refinement.

Regarding the three Hypotheses stated in section 8.1 the conclusion reached based on the research questions is as follows:

H1: The generated text is accurate

This hypothesis is rejected since the participants found the conclusion generated to be insufficiently accurate despite considering the other information sufficiently accurate

H2: The generated text is readable and understandable.

This hypothesis is confirmed due to the positive results of questions 2 & 3.

H3: The generated natural language summary of fund data makes a positive difference as compared to a graph-only presentation of the processed data, in the context of audits.

This hypothesis is confirmed due to the positive results in question 5

8.9 Final Improvements

8.9.1 Summary of Changes

Following the evaluation changes were made to the design which are detailed in Table 20

Table 20: A summary of possible changes as well as whether they were implemented in the design. Items shaded blue are already discussed in the previous section whereas those in orange are further discussed in this section.

ID	Task Area	Evaluation Objective	Potential Change to be implemented	Implemented?	Follow up evaluation required
1	Content Determination	Accuracy (conclusion)	Alteration to conclusion tie-breaker algorithm	No, accuracy gain not certain	Large scale user evaluation in combination with Machine Learning and more possible metrics
2	Content Determination	Accuracy (Other Information)	Static number of alternative trend summaries to a dynamic number based on the size of the deviations	Yes	No, change not significant enough, overall summary not significantly affected
3	Sentence planning	Readability	More variability in text summaries	No, preference of participants not clear enough	N/A
4	Lexicalisation	Synonym Choices	Replace "metric" with "characteristic"	Yes	No, clear preference and change is highly unlikely to have "side-effects"
5	Lexicalisation	Synonym Choices	Replace "Synthetic Index" with "Benchmark"	Yes	No, clear preference and change is highly unlikely to have "side-effects"

6	Content Determination & Sentence planning	Other Suggestions	Adding benchmark growth to sentence around fund growth	Yes	No, change not significant enough, overall summary not significantly affected
7	Sentence planning	Other Suggestions	adding items to be followed up on by the audit team	Yes	No, change not significant enough, overall summary not significantly affected
8	Sentence planning	Other Suggestions	changing conclusion structure	No, change too large and only one evaluator behind it	N/A
9	Sentence planning	Other Suggestions	adding additional conclusion	No, change too large and only one evaluator behind it	N/A

The logic behind some of the changes and their implementation (or lack of implementation) has already been covered in the previous section. The remainder are discussed below with reference to Figure 35 which shows a comparison output before and after the change with the differences highlighted.

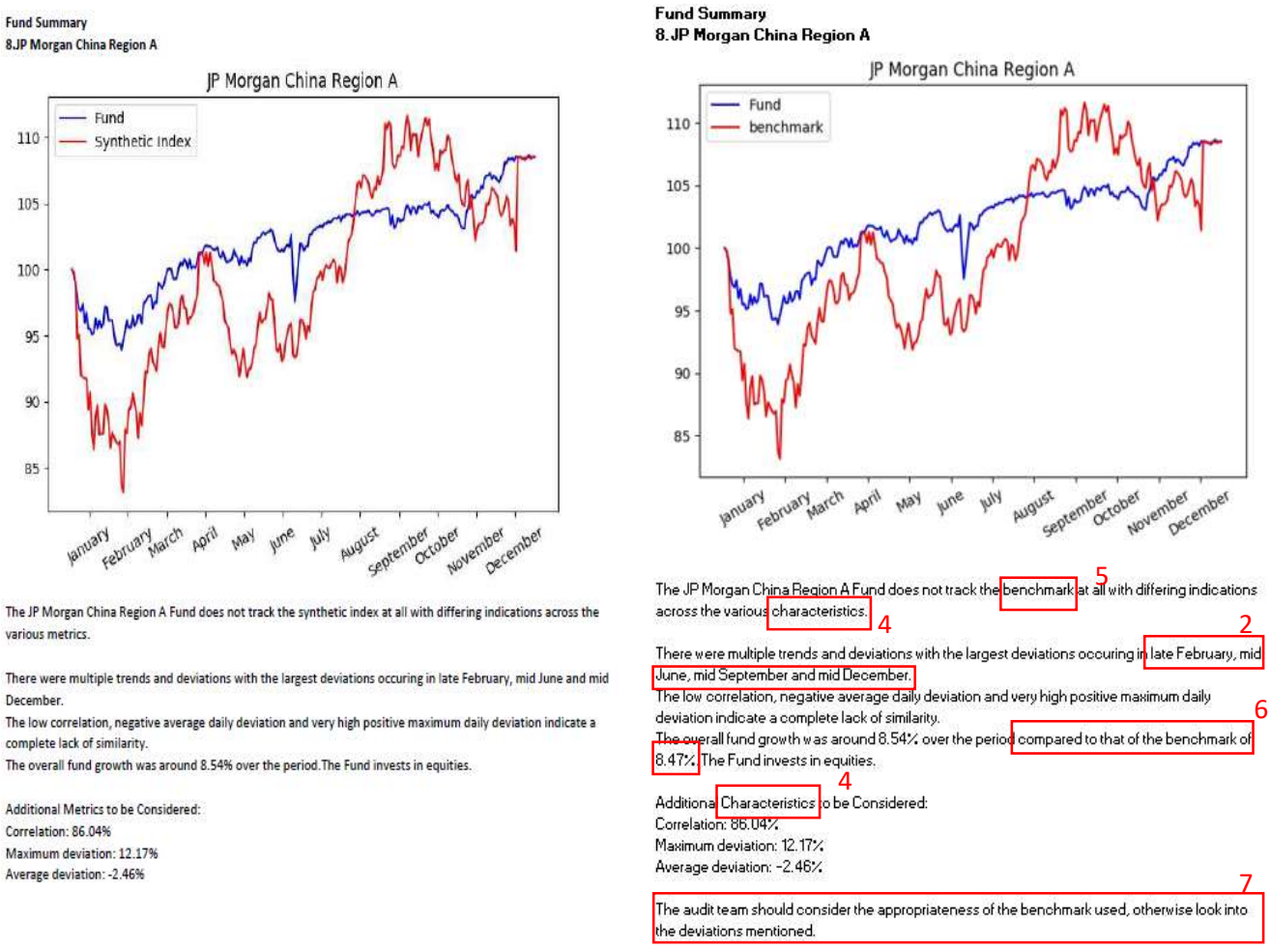


Figure 35: An example comparison of pre- and post-evaluation outputs with the changes to the summary shown on the right in red boxes.

8.9.2 Change 2: From static to dynamic alternative trends summaries

Recalling Section 8.7.4 regarding trends which were not identified by the design, the revised algorithm continues adding trends to the summary until their maximum deviation is no longer greater than 60% of the max deviation across the period. This percentage was selected by starting at an arbitrary percentage and adjusting until the deviations identified as missing by the participants were included in the summary. The result is a more comprehensive description of the major trends where a detailed analysis is not possible. In addition the result is more natural since it is not predefined as a specific number of trends to mention. The largest number of trends described in the test funds was 6 for fund 7. While this could be considered lengthy the trade-off is considered acceptable as there is less risk of ignoring key trends. The revised pseudocode is shown in Figure 36.

```
1: procedure ShortenedSummarySelection
2:     for Trends in TrendSummary do
3:         Trend.maxDD ← Absolute(Trend.maxDD)
4:     LargestDD ← TrendWithHighestMaxDD(TrendSummary)
5:     TrendCutoff ← LargestDD * 0.6
6:     while LargestDD > TrendCutoff do
7:         Top DDs.add(LargestDD)
8:         TrendSummary.remove(LargestDD)
9:         LargestDD ← TrendWithHighestMaxDD(TrendSummary)
```

Figure 36: The pseudocode presented in Figure 24 where the code was adjusted to consider all deviations greater than 60% of the MaxDD. Lines 4 onwards have been adjusted.

8.9.3 Change 7: Adding follow up action for audit team

At a simple level all the trends identified in the summary could have been listed at the bottom of the summary. This was considered redundant as the information has already been displayed in the body of the summary. Instead the choice was made to simply refer the reader to the aspects they need to consider which also takes into the account the overall conclusion. The case table for determining the follow up is shown in Table 21 which includes the rationale for the changes. All output text shown below is canned.

Table 21: The complete case table for selecting the follow up suggestion as well as the rationale for the suggestions.

Case	Result	Rationale
Conclusion "tracks the benchmark" and specific situation "isolated spike" identified?	"The audit team should consider enquiring as to the reason for the isolated spike before concluding."	While there is unlikely to be any issue with the fund development, it can expose the audit team to scrutiny if a spike is ignored completely
Conclusion "does not track effectively"	"The audit team should investigate the deviations mentioned above before concluding."	The more significant deviations (which are causes/symptoms of the conclusion) need to be investigated before the team can conclude.
Conclusion "does not track at all"	"The audit team should consider the appropriateness of the benchmark used, otherwise look into the deviations mentioned."	Given how poor the time series match, it is possible that the benchmark used is not in fact appropriate for the fund, however if this is not the case the deviations should be followed up on in order to conclude.
Default	Null	This would be an instance where the result is satisfactory and no unusual circumstances are identified (since only an isolated spike can occur in a satisfactory result). No follow up is therefore required by the audit in order to conclude.

This change provides some more guidance and provides a clear path to concluding on the information provided as requested by some of the participants. This is also consistent with the authors understanding of the auditing process, but which had not been previously considered.

8.9.4 Need for Further Evaluation?

Based on the changes made above and as described in Table 21, a new evaluation cycle is not needed to meet the stated objectives in Section 8.1. In some cases the changes are clear-cut (synonym replacement), and there is no other impact on the summary that could have an unintended effect on the accuracy, understandability or readability of the summaries. For others, there is a potential trade-off of understandability/accuracy and readability (changes 2 & 6) however the changes are not considered significant enough to warrant another evaluation. Lastly, regarding number 7, the additional sentence could potentially be considered significant. Despite this, the change is highly logical and necessary in the audit domain where follow up action would be required by the audit team in the cases identified. Additionally, since the question of its inclusion was not specifically asked (as its inclusion was not considered prior to the evaluation), the fact that two participants noticed its omission is considered significant. Considering this, the inclusion of this follow up is considered necessary to achieve the design's objective. Considering the participants potential assessment of the wording, the text output

added is not “strict” in wording nor is it highly specific on what exact deviations to consider and how to follow up. It is therefore general enough that the participants would not have grounds to reject such an improvement.

9 Discussion

As a starting point for the overall discussion, the evaluation method is first reflected on as well as the outcome thereof. In addition the design is compared to the state of the art in the auditing domain as well as in general. Lastly, the possibility for future work and research is presented.

9.1 Reflection on Evaluation Method

Regarding responses, the number and quality of responses was high. All expert participants provided responses with only some minor omissions. Qualitative feedback (e.g., Question series D) was also provided with all except one participant providing general improvement points in Question 19.

In terms of consistency on responses, among the participants different aspect had different levels of agreement. For example, while there was not much division in determining the correct conclusion (Question Series A), there was significant division in determining alternative word choice (Question 16-18). While this is not necessarily bad it does make it harder to identify improvements, since there is effectively no way to please all or even most of the participants. This means that any changes where there is sufficient division would require sound logical backing in order to implement.

In addition, there were some minor differences in understanding of the questions by the participants as shown by the same suggestions being made in different qualitative questions e.g., Participant 5 indicating that they did not understand a metric in Question Series D which deals specifically with accuracy . The suggestions themselves remain valid in spite of this.

With regard to the number of participants, 7 is considered sufficient for the purposes of the evaluation, as these are experts in the field. This fulfilled the requirement for concluding on the three hypotheses.

Following the evaluation, the design has been improved through the addition of useful information, better selection of synonyms and a minor change to the trend summary logic.

9.2 Comparison with other works

In comparing the design to the state of the art, Figure 37 provides an overview of the various stages involved in the design. This consists of (Reiter & Dale, 1997)'s NLG stages as well as an additional (not strictly NLG) data-to-information stage (discussed in Section 3.1). On an NLG basis, the design is straightforward without significant advancements in sentence planning and linguistic realisation. It utilises an existing solution (SimpleNLG) to effectively create understandable and readable text, according to industry experts.

Regarding text planning, the system introduces a novel way to handle the description of "actor" and passive time series in performing a comparison (Section 6.4). Additional, smaller, innovations were made without reference to existing literature (such as the descriptions of months based on whole number and decimal separation, Section 7.2.1). Since these are not complicated techniques, it is possible that they have been previously implemented but not widely noted in the existing literature.

Considering the data-to-information stage, the application of fuzzy logic to deciding on overall conclusions is novel as far as the author is aware. While (Kacprzyk & Wilbik, 2009)'s design (the closest comparable work found) also uses fuzzy logic in time series comparison, their application is to determine the extent to which human descriptions of two trends are similar. The current design tackles the issue in a different manner, using a fuzzy combination of key metrics to form a conclusion. In addition, this design has been evaluated by domain experts whereas (Kacprzyk & Wilbik, 2009)'s evaluation assumed similarity between two times series and then evaluated the numeric results of their design.

Additionally the comparison of time series based on a derivative time series representing the difference between the original two (refer Section 6). Inspiration is drawn from (Kacprzyk, Wilbik & Zadrozny, 2006) who use the cone graph linear approximation technique (Section 6.1) on the original times series to summarise the trends linguistically. In the current design however the same technique is performed on the derivative time series to allow for a richer description of the deviations between the trends.

In summary, when compared with the state of the art in NLG, the current design provides novel solutions in the earlier NLG stages (as well as pre-NLG stages) with less innovation in the later stages.

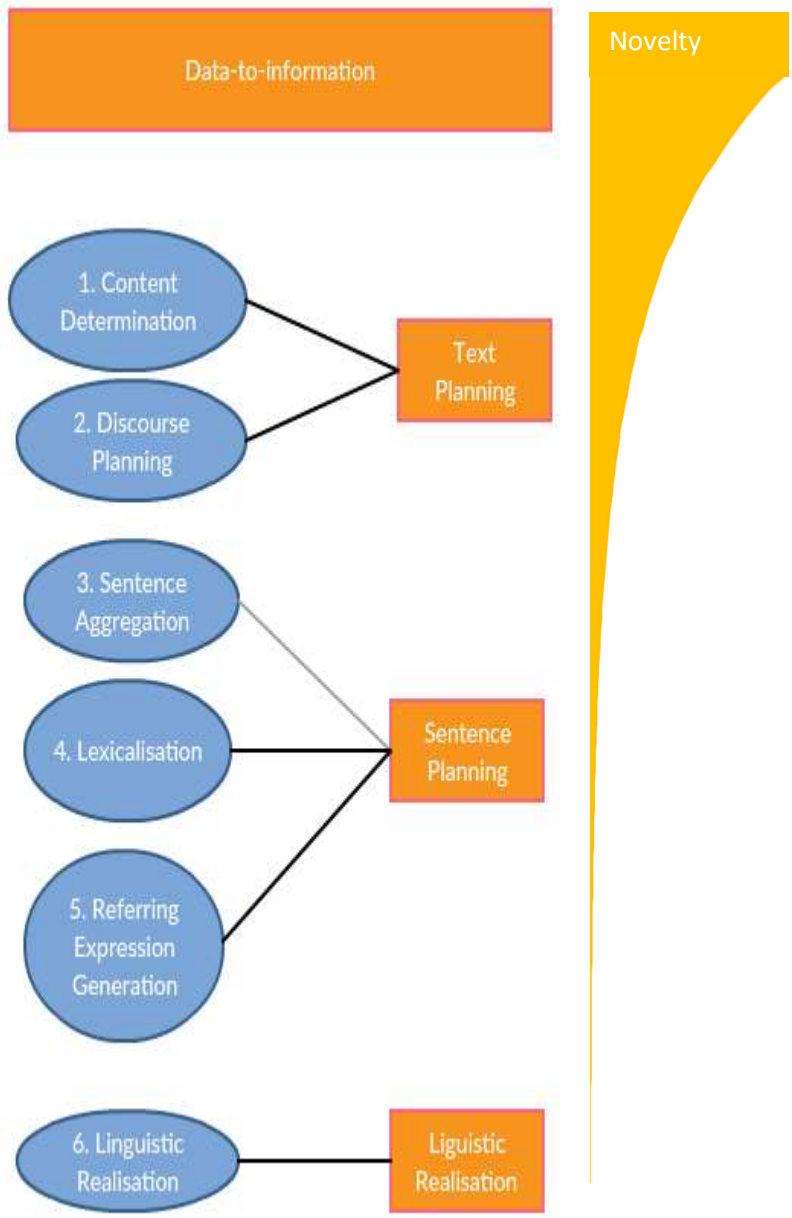


Figure 37: A repetition of Figure 2 including an additional data-to-information stage as well as an indication of the novelty of the current design compared to the state of the art.

Comparing to other works in the auditing domain, the current design stands apart as there are no similar works as far as the author is aware. Many existing techniques, not novel in the NLG space, have been researched and adapted to the problem area of auditing investment funds. More generally, the combination of applying decision making in moving from data to information has not been performed in the auditing domain previously.

9.3 Further works

During the design phase, additional avenues of research were identified outside of the scope of the current research. One such area was the incorporation of the nature of the fund i.e., the type of assets making up the fund, into the conclusion determination task. For example a passive fund would not be expected to change as much in composition over the course of the year. This means that a much higher correlation would be expected with the index, compared to an active fund whose composition may change throughout the year since it makes trades independent of the benchmark. Such functionality would require a fully functional user friendly system which is not in the scope of this research and is therefore considered a future research point.

Another potential research topic is the refinement of the trend detection method used in the time series comparison. A large range of possibilities were noted with different strengths and weaknesses. Separate research could be performed to determine the ideal technique to apply to the problem of identifying trends in time series.

Additional areas of research were identified based on participant evaluation. One of these was the failure of the design to achieve the target accuracy score. To solve this issue, One possibility would be to obtain a sample of funds with expert conclusions and use these in a regression where multiple metrics are tested for their predictive power (this is a simplification however due to the fuzzy logic used to process the metrics). This could be used to fine tune the metrics included in the conclusion determination. Another possibility for further research is to test whether the experts are swayed by graphical factors not incorporated in the decision-making of the design. Identifying and understanding these factors would allow for adjustments to be made to the logic to improve accuracy. Alternatively it may provide evidence that the human experts are subject to errors in perception, and that the system is providing the more consistent conclusion.

The final significant avenue is the determination of reasons for deviations identified by the system i.e., not only describing deviations but providing potential reasons. This would require the inclusion of additional data sources. One such possibility would be the inclusion of financial news articles in the country/industry applicable to the fund being compared. Natural Language Processing could be used to extract key information from these articles, which could be applied to deviations noted in the time series to determine if the key information could explain the deviation.

10 Conclusion

Expert evaluation showed that the text produced by the design is both readable and understandable (confirmation of H2). Furthermore there is almost unanimous agreement that the design makes a positive impact on both audit quality and confirming H3. This indicates that the system does indeed make a positive impact on the auditing domain. The design does however, still need improvement to its overall accuracy (rejection of H1).

In designing this system, advancements were made in data-to-information processing and to a lesser extent, text planning within the classic NLG framework which can be applied to the field as a whole. Additionally, this research provides an initial application of multiple NLG techniques into the domain of auditing, which had not previously been done.

Further research has also been proposed to further refine the trend identification through additional literature review and testing. In addition, the accuracy of the design can be further improved through regression analysis, in order to improve the selection of metrics. Another possibility is to investigate whether the participants are affected by graphical factors, and conclude whether this improves or worsens the accuracy of their results. Finally, the design could be expanded to offer possible explanations for trends and deviations identified, incorporating other data sources, such as news feeds, to further improve the designs ability to support auditors.

11 Reference List

Accountingverse 2017. *Big 4 Accounting Firms.* Available: http://www.accountingverse.com/articles/big-4-accounting-firms.html; [2017].

Adeyanju, I. 2012. Generating weather forecast texts with case based reasoning. *International Journal of Computer Applications.* 45:35-40.

Bateman, J. 2016. *KPML One-Point Access Page.* Available: http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/README.html; [2017].

Castillo-Ortega, R., Mann, N. & Sánchez, D.2011. Linguistic local change comparison of time series. *Fuzzy Systems (FUZZ), 2011 IEEE International Conference On.* IEEE. 2909.

Dale, R., Geldof, S. & Prost, J.2003. CORAL: Using Natural Language Generation for Navigational Assistance. *ACSC '03 Proceedings of the 26th Australasian Computer Science Conference.* 2003. M. Oudshoorn, Ed. Darlinghurst, Australia: Australian Computer Society, Inc. 35.

Dale, R. & Mellish, C.1998. Towards evaluation in natural language generation. *In Proceedings of First International Conference on Language Resources and Evaluation.*

Dan, J., Shi, W., Dong, F. & Hirota, K. 2013. Piecewise trend approximation: a ratio-based time series representation. *Abstract and Applied Analysis.* 2013:Article ID 603629.

Debled-Rennesson, I., Tabbone, S. & Wendling, L.2004. Fast polygonal approximation of digital curves. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference On.* IEEE. 465.

Fu, T. 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence.* 24(1):164-181.

Gatt, A. & Reiter, E.2009. SimpleNLG: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009).* 2009. Association for Computational Linguistics. 90.

Goldberg, E., Driedger, N. & Kittredge, R.I. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert.* 9(2):45-53.

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering.* 9(3):90-95.

Kacprzyk, J. & Wilbik, A.2009. Using Fuzzy Linguistic Summaries for the Comparison of Time Series: an application to the analysis of investment fund quotations. *IFSA/EUSFLAT Conf.* 1321.

Kacprzyk, J. & Wilbik, A.2010. A comprehensive comparison of time series described by linguistic summaries and its application to the comparison of performance of a mutual fund and its benchmark. *Fuzzy Systems (FUZZ), 2010 IEEE International Conference On.* IEEE. 1.

Kacprzyk, J., Wilbik, A. & Zadrozny, S.2006. Linguistic summarization of trends: a fuzzy logic based approach. *Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems.* 2166.

Keogh, E.J. & Pazzani, M.J. 1998. An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. *Kdd.* 98(1):239-243.

Knight, K. & Hatzivassiloglou, V.1995. Two-level, many-paths generation. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics.* 1995. Association for Computational Linguistics. 252.

Lewis, D.D., Yang, Y., Rose, T.G. & Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research.* 5(Apr):361-397.

Lin, J., Keogh, E., Lonardi, S. & Chiu, B.2003. A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.* ACM. 2.

Moraes, P., McCoy, K. & Carberry, S.2016. Enabling text readability awareness during the micro planning phase of NLG applications1. *The 9th International Natural Language Generation Conference.* 2016. 121.

Moraes, P., Sina, G., McCoy, K. & Carberry, S.2014. Evaluating the accessibility of line graphs through textual summaries for visually impaired users. *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility.* ACM. 83.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y. & Sykes, C. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence.* 173:789-816.

Ramos-Soto, A., Vazquez-Barreiros, B., Bugarín, A., Gewerc, A. & Barro, S. 2017. Evaluation of a Data-To-Text System for Verbalizing a Learning Analytics Dashboard. *International Journal of Intelligent Systems.* 32(2):177-193.

Ramos-Soto, A., Bugarín, A.J., Barro, S. & Taboada, J. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems.* 23(1):44-57.

Reiter, E. & Dale, R. 1997. Building Applied Natural Language Generation Systems. *Natural Language Engineering.* 1:57-87.

Reiter, E.2007. An architecture for data-to-text systems. *Proceedings of the Eleventh European Workshop on Natural Language Generation.* Association for Computational Linguistics. 97.

Reiter, E. & Sripada, S.2002. Should corpora texts be gold standards for NLG. *Proceedings of INLG.* 97.

SkilsYouNeed 2018. *Percentage change | Increase and Decrease.* Available: https://www.skillsyouneed.com/num/percent-change.html; [2018].

Sklansky, J. & Gonzalez, V. 1980. Fast polygonal approximation of digitized curves. *Pattern Recognition.* 12(5):327-331.

Smiley, C., Plachouras, V., Schilder, F., Bretz, H., Leidner, J.L. & Song, D.2016. When to Plummet and When to Soar: Corpus Based Verb Selection for Natural Language Generation. *The 9th International Natural Language Generation Conference.* 5 September 2016. 36.

Song, Q. & Chissom, B.S. 1993. Fuzzy time series and its models. *Fuzzy Sets and Systems.* 54(3):269-277.

Statistics How To 2018. *About Normalized Data.* Available: http://www.statisticshowto.com/normalized/; [2018].

Van Deemter, K., Theune, M. & Krahmer, E. 2005. Real versus template-based natural language generation: A false opposition? *Computational Linguistics.* 31(1):15-24.

Yin, P. 2004. A discrete particle swarm algorithm for optimal polygonal approximation of digital curves. *Journal of Visual Communication and Image Representation.* 15(2):241-260.

Zadeh, L.A. 1973. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. *IEEE Transactions on Systems, Man, and Cybernetics.* SMC-3(1):28-44. DOI:10.1109/TSMC.1973.5408575.

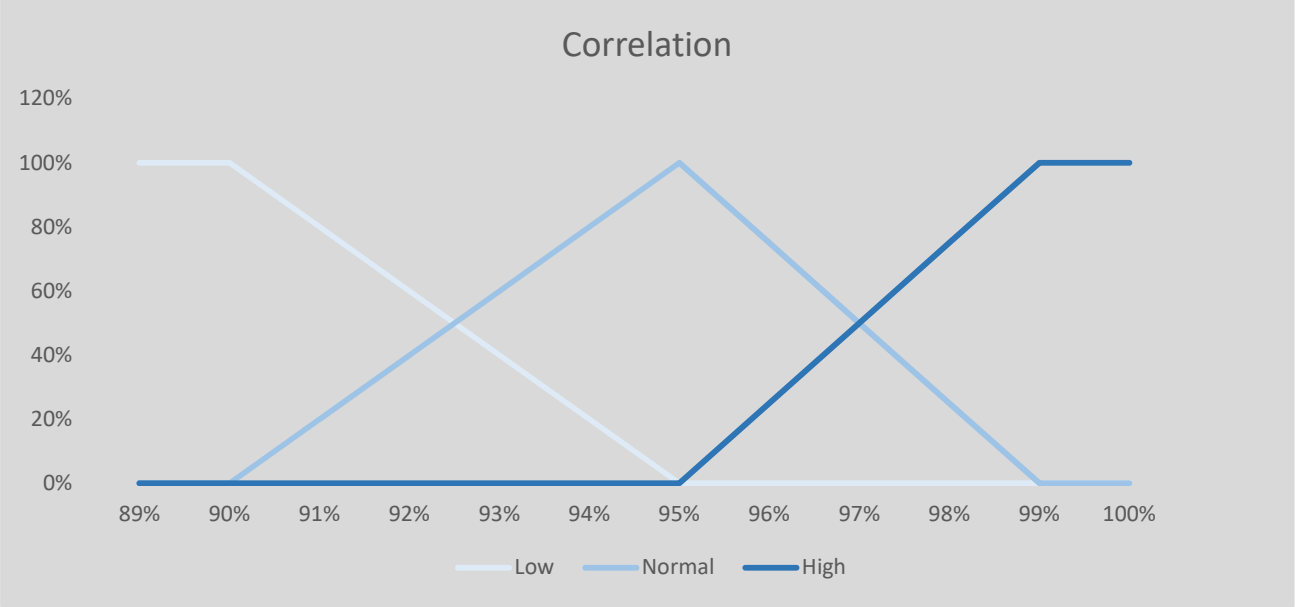
Zadeh, L.A. 1997. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems.* 90(2):111-127.

Zadeh, L.A.2002. A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium.* IEEE. 2.

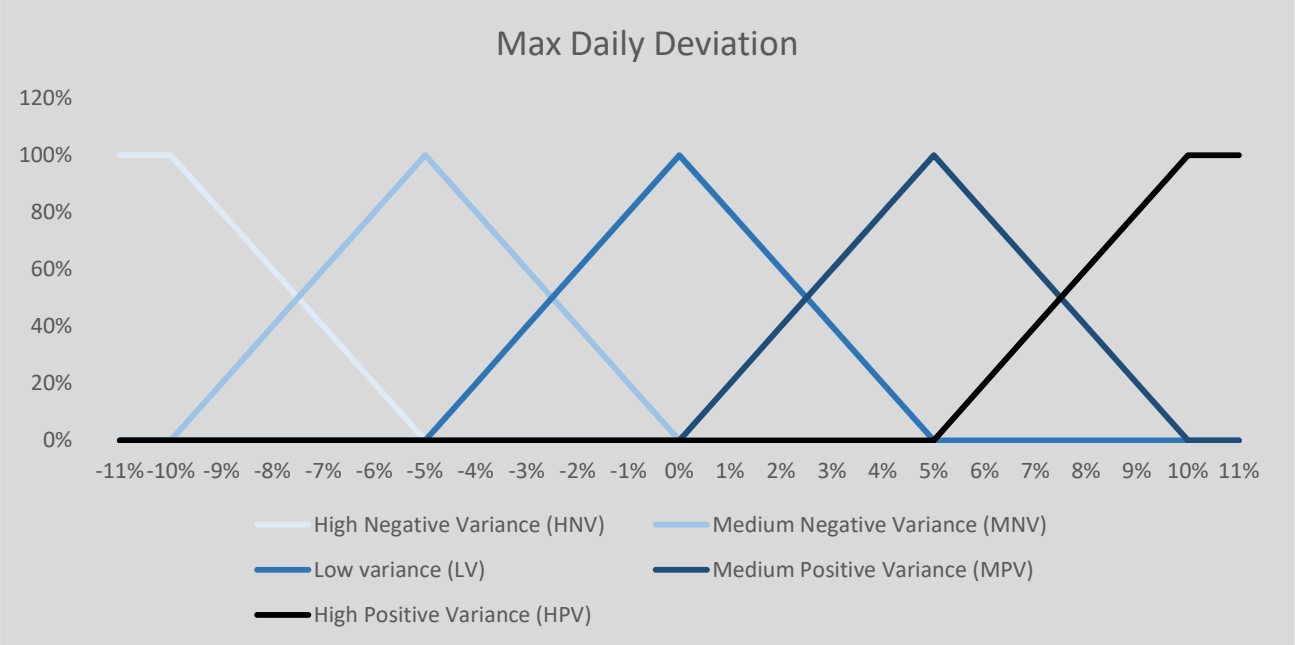
Zadeh, L.A. 1965. *Fuzzy sets.* Available: http://www.sciencedirect.com/science/article/pii/S001999586590241X .

12 Appendix A: System Functions/Mappings

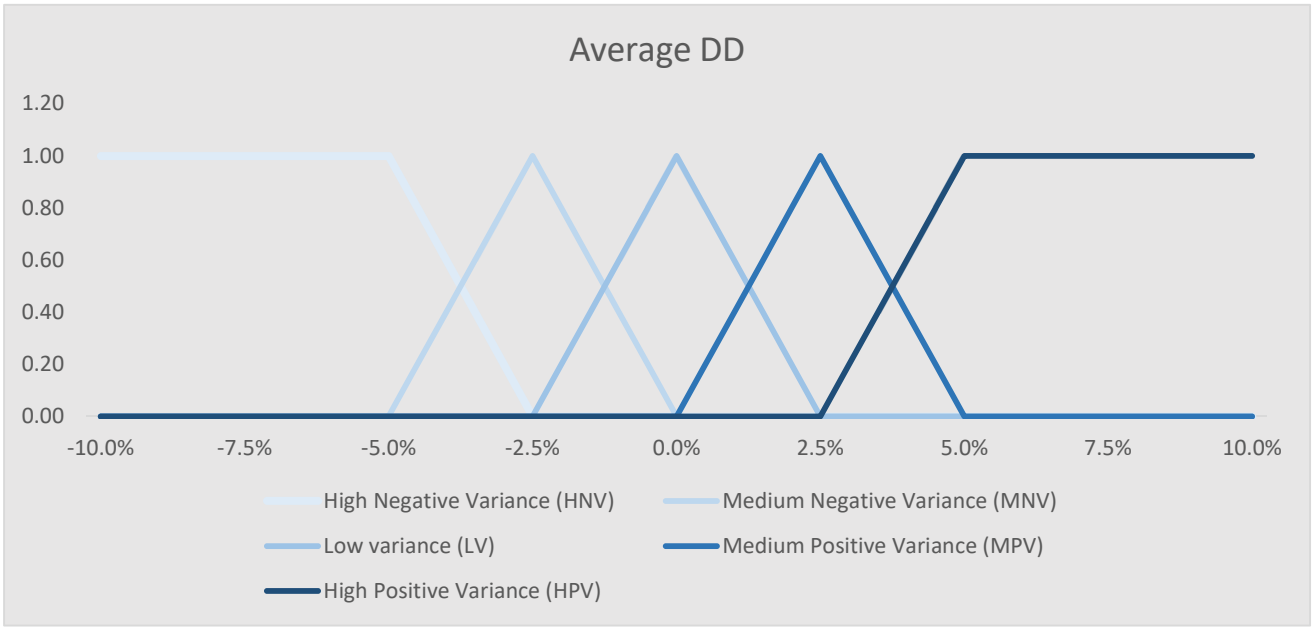
12.1 Fuzzy Membership Functions

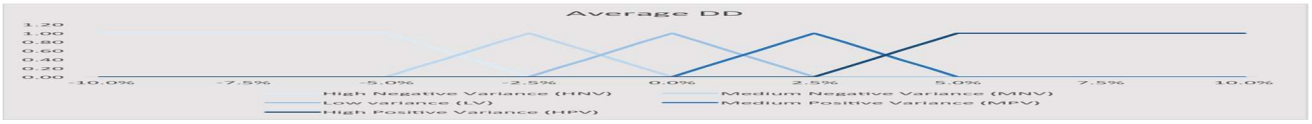


X0	X1	X2	X3	X4	X5	X6
-10%	0%	90%	95%	99%	100%	102%



X0	X1	X2	X3	X4	X5	X6	X7	X8
-101%	-100%	-10%	-5%	0%	5%	10%	100%	101%


Average DD
1.20
1.00
0.80
0.60
0.40
0.20
0.00
-10.0%
-7.5%
-5.0%
-2.5%
0.0%
2.5%
5.0%
7.5%
10.0%
High Negative Variance (HNV)
Medium Negative Variance (MNV)
Low variance (LV)
Medium Positive Variance (MPV)
High Positive Variance (HPV)



12.2 Specific Situation Mapping

Maximum Daily Deviation	Correlation	Average Daily Deviation	Specific Situation
MPV	**high**	**MPV**	**likelihood of distributions from Fund**
MPV	**high**	**HPV**	**likelihood of distributions from Fund**
HPV	**high**	**MPV**	**likelihood of distributions from Fund**
HPV	**high**	**HPV**	**likelihood of distributions from Fund**
MNV	**high**	**MNV**	**potential for artificial inflation**
MNV	**high**	**HNV**	**potential for artificial inflation**
HNV	**high**	**MNV**	**potential for artificial inflation**
HNV	**high**	**HNV**	**potential for artificial inflation**
MPV	**high**	**LV**	**isolated spike**
HPV	**high**	**LV**	**isolated spike**
MPV	**high**	**MNV**	**isolated spike**
MPV	**high**	**HNV**	**inconsistency**
HPV	**high**	**MNV**	**inconsistency**
HPV	**high**	**HNV**	**inconsistency**
HPV	**normal**	**HNV**	**inconsistency**
LV	**high**	**HPV**	**inconsistency**
LV	**high**	**HNV**	**inconsistency**
LV	**normal**	**HPV**	**inconsistency**
LV	**normal**	**HNV**	**inconsistency**
LV	**low**	**HPV**	**inconsistency**
LV	**low**	**HNV**	**inconsistency**
MNV	**high**	**HPV**	**inconsistency**
HNV	**high**	**MPV**	**inconsistency**
HNV	**high**	**HPV**	**inconsistency**
HNV	**normal**	**HPV**	**inconsistency**
MPV	**normal**	**HNV**	**complete lack of similarity**
MPV	**low**	**HNV**	**complete lack of similarity**
HPV	**normal**	**MNV**	**complete lack of similarity**
HPV	**low**	**HNV**	**complete lack of similarity**
HPV	**low**	**MNV**	**complete lack of similarity**
HNV	**normal**	**MPV**	**complete lack of similarity**
HNV	**low**	**MPV**	**complete lack of similarity**
HNV	**low**	**HPV**	**complete lack of similarity**
MNV	**normal**	**HPV**	**complete lack of similarity**
MNV	**low**	**HPV**	**complete lack of similarity**

Function	Relevant Variable in information structure in Section 7.1	Variable	Explanation
Flat	TrendSummary	start	beginning of the flat trend
		end	end of the flat trend
Trends	TrendSummary	Months	months in which the trend occurred
		Actor	either the Fund or the synthetic index depending on the results of the results of Section 6.4
		obj	either the Fund or the synthetic index depending on the results of the results of Section 6.4
		movement	either "grow" or "drop" depending on the results of Section 6.4
		modifier	either "ahead of" or "behind" depending on the results of Section 6.4
		sig	whether or not the gradient is "sharp"
		spike	whether or not the trend is a spike (Section 6.3)
		isF	whether or not the trend is flat
		realNum	the number of the trend that is mentioned (.e.g., if two trends have already been realised, then this trend would be realNum = 3)
Specific Situations	SpecialSituationSummary	factor1	partially realised metric 1
		factor2	partially realised metric 2
		factor3	partially realised metric 3
		specSit	Special situation as determined in Specific Situations Mapping
Supplementary Realisation	fundGrowth, PrimaryInvestment, SecondaryInvestment	primInstrs	all but one primary instrument
		primLastInstr	last primary instrument (joined to the others with "and")
		secInstrs	All but one seconday instrument
		secLastInstr	last secondary instrument (joined to the other with "and")

Alternative Trend Summary	TrendSummary	month1	first month description in which large deviation occurred
		month2	Second month description in which large deviation occurred
		month3	Third month description in which large deviation occurred

13 Appendix B: Questionnaire

13.1 Evaluation for Fund Comparison Natural Language Generator

Evaluator Number: _____

The aim of this evaluation is to determine the accuracy, understandability and readability of the textual summaries generated. In addition it aims to determine if the textual summaries provided add value to the audit, either through improving audit efficiency or audit quality. Note that in some cases data may have been modified or benchmarks changed to simulate deviations for evaluation purposes.

The following questions relate to each individual fund summary provided to you in “Fund Summaries.pdf”. There are 10 summaries each for a different fund. Please select your answer for each question.

NB: Please ensure that you have understood and signed the consent form prior to starting this evaluation.

1. Fidelity Growth Company
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

2. Vanguard Balanced
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

3. JP Morgan Balanced
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

4. Vanguard High Yield
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

5. Vanguard Wellington
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

6. Vanguard S&P 500 Growth
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

7. JP Morgan Hedged Equity
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

8. JP Morgan China Region A
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

9. Mirae Emerging Markets
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

10. Modified Illustration
    a. How would you rate the accuracy of the conclusion “e.g. The Vanguard Balanced Fund tracks the benchmark effectively”

        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    b. If you answered fairly inaccurate or highly inaccurate to question a. what would your conclusion be?

        - The fund tracks the benchmark effectively
        - The Fund does not track the benchmark effectively
        - The Fund does not track the benchmark at all

    c. How would you rate the accuracy of the other information provided?
        - Highly accurate
        - Fairly accurate
        - Fairly inaccurate
        - Highly inaccurate

    d. If you answered fairly inaccurate or highly inaccurate to question c. please explain below what you found inaccurate

    e. How much of the information did you find understandable?
        - I understood all the information presented
        - I understood most of the information presented
        - I did not understand most of the information presented
        - I did not understand any of the information presented

The below questions relate to the summaries in general.

11. Please indicate which pieces of information you did not understand in the summaries provided

12. Please rate the repetitiveness of the text across the different summaries
    - Highly repetitive
    - Moderately repetitive
    - Not repetitive

13. There are multiple ways of putting into words the same thing. For instance

"The high correlation, low average daily deviation and positive maximum daily deviation indicate an isolated spike."

The text could also read on some of the summaries

"An isolated spike is indicated by the high correlation, low average daily deviation and positive maximum daily deviation"

Would you prefer the same structure of the sentences across the summaries, keep the summaries as is, or more of this sort of variation across the summaries in describing the information?

- More Variability
- No change
- Less Variability

Note: the purpose of this question is to gauge the evaluator's assessment of the trade in variability. As variability increases the text would appear more natural, human and potentially more pleasant to read. At the same time though the changing in position of critical information (in this case "isolated spike") could cause frustration for the user as they try to scan over the text and pick out the important information quickly

14. Do you think that the textual summary impacts audit efficiency compared to the basic summary?
    4. Improves audit efficiency
    5. Make no difference to audit efficiency
    6. Worsens audit efficiency

15. Do you think that the textual summary impacts audit quality compared to the basic summary?
    4. Improves audit quality
    5. Make no difference to audit quality

6. Worsens audit quality

16. In place of the verb “track” used in the conclusion part of the summary, would you prefer to see:
    1. Follows
    2. Mirrors
    3. Is similar to
    4. Other: _____

    e.g. The Fidelity Growth Company Fund does not **mirror** the benchmark effectively

17. In place of the term “metric” used in the conclusion part of the summary, would you prefer to see:
    a. Measure
    b. Characteristic
    c. Indicator
    d. Other: _____

    e.g. Additional **Measures** to be Considered

18. In place of the term “synthetic index” used throughout the summary, would you prefer to see:
    a. Benchmark
    b. Artificial index
    c. Other: _____

    e.g. The Fidelity Growth Company Fund does not track the **benchmark** effectively

19. Do you have any further suggestions for improvement or comments?

14 Appendix C: Funds Covered in Assessment

Fund Number	Fund Name	Source Price Data	Indices Used	Index Source Data	Period
1	Fidelity Growth Company	Yahoo Finance	Russell 3000 Growth Index	Yahoo Finance	01/01/2016-31/12/2016
2	Vanguard Balanced	Yahoo Finance	40% S&P U.S. Aggregate Bond Index & 60% S&P 500 Index	us.spindices.com	01/01/2016-31/12/2016
3	JP Morgan Balanced	Yahoo Finance	50% S&P U.S. Aggregate Bond Index & 50% S&P 500 Index	us.spindices.com	01/01/2016-31/12/2016
4	Vanguard High Yield	Yahoo Finance	S&P 500 High Yield Corporate Bond Index	us.spindices.com	01/01/2016-31/12/2016
5	Vanguard Wellington	Yahoo Finance	35% S&P U.S. High Yield Low Volatility Corporate Bond Index & 65% S&P 500 Index	us.spindices.com	01/01/2016-31/12/2016
6	Vanguard S&P 500 Growth	Yahoo Finance	S&P 500 Growth	us.spindices.com	01/01/2016-31/12/2016
7	JP Morgan Hedged Equity	Yahoo Finance	S&P 500 Growth	us.spindices.com	01/01/2016-31/12/2016
8	JP Morgan China Region A	Yahoo Finance	S&P China BMI (US Dollar)	us.spindices.com	01/01/2016-31/12/2016
9	Mirae Emerging Markets	Yahoo Finance	S&P BSE SENSEX	us.spindices.com	01/01/2016-31/12/2016
10	Modified illustration (Based on Vanguard S&P 500 Growth)	Yahoo Finance	S&P 500 Growth	us.spindices.com	01/01/2016-31/12/2016